EFFECTS OF DISRUPTING THE INTERACTION OF MURINE LEUKEMIA VIRUS (MLV) WITH HOST CHROMATIN-BINDING PROTEIN ON TUMORIGENESIS IN *MYC/RUNX2* MOUSE MODEL AND RECOGNITION OF CHROMATIN

By

LORENZ M. LOYOLA

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Biochemistry

written under the direction of

Monica J. Roth

and approved by

____

____

____

____

New Brunswick, New Jersey

MAY, 2019

ABSTRACT OF THE DISSERTATION

Effects of disrupting the interaction of murine leukemia virus (MLV) with host chromatin-binding protein on tumorigenesis in *MYC/Runx2* mouse model and recognition of chromatin

by LORENZ M. LOYOLA

Dissertation Director:
Monica J. Roth

Murine leukemia virus (MLV) integrase (IN) lacking the C-terminal tail peptide (TP) lose the interaction with the host bromodomain and extraterminal (BET) proteins and decrease their integration preference at promoter/enhancers and transcriptional start sites and CpG islands. MLV lacking the IN TP by altering the open reading frame were infected into a tumorigenesis mice model (*MYC/Runx2)* to observe integration patterns and phenotypic effects. Viral passage resulted in the restoration the TP onto IN through small deletions. Mice infected with different modified MLV lacking the IN TP^{-} coding sequence (TP^{-}), showed an improved median survival by 10 days compared to wildtype (WT) MLV infection. Recombination with polytropic endogenous retrovirus (ERV), *Pmv-20,* were identified in seven mice, displaying both fast and slow tumorigenesis. Next generation sequencing of tumors showed an infected mouse (TP^{-}16) without observed recombination with ERVs with less integrations at TSS and CpG islands, compared to the mean integrations of WT tumors. This mouse also has less integrations at Brd4 and BET associated histone modifications (H3K4me1/3) +/- 1 kb from ChIP-seq peaks. However, this mouse succumbed to the tumor rapidly (34 days). Analysis of the four of the top copy number integrants of the TP^{-}16 tumor revealed their proximity to known MLV common insertion sites genes (*Hdac6, Ccnd1, Rasgrp1),* maintaining their MLV IN TP^{-} genotype. Furthermore, mapping integrations in K562 cells revealed the preference of MLV IN TP- insertions within chromatin profile states associated with heterochromatin and weakly transcribed regions. A decreased number of integrations were observed at histone marks associated with BET proteins (H3K4me1/2/3, and H3K27Ac). The results highlight the strong selection within the mouse to maintain the full-length IN protein. MLV IN TP$^-$ showed a decreased overall rate of tumorigenesis compared to WT virus in the *MYC/Runx2 model*. However, MLV integrations, in the absence of the influence of BET proteins, can still occur at regions of oncogenesis driver genes, either stochastically or through trans-complementation by functional endogenous Gag-Pol. Thus, the modified MLV virus can be a safer vector than the wildtype virus, but it still maintains the oncogenic potential. This study provides new insights on how to improve the safety of MLV retroviral vectors.

Acknowledgment

I would like to express my sincere gratitude to my advisor Dr. Monica Roth and the rest of the members of the laboratory throughout the years for the continuous support of my research work in my Ph.D. study. I truly value Dr. Roth's immense knowledge and patience in guiding me through my research work. Her guidance has helped me through all my shortcomings and scientific endeavors.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Lisa Denzin, Dr. Gaetano Montelione, and Dr. Arnold Rabson, for their encouragement, and insightful comments. Special thanks to Dr. Montelione for the guidance and advice he gave me during my second qualifying exam. I would also like to thank Dr. Nancy Walworth and Dr. Janet Alder for being valuable mentor from the beginning of my Ph.D. program to my involvement in Molecular Biosciences Graduate Student Organization and Rutges iJOBS.

I'd like to acknowledge our research collaborators, Dr. Vasudevan Achutan, Dr. James Neil, Dr. Alan Engelman, Dr. Christine Kozak, and some of the members of their research group. I appreciate the knowledge and resources they have contributed for the completion of this research.

I would also like to extend my thanks to the department administrative team, technicians and past members of the laboratory for helping me navigate the lab and teaching me experiments integral to my research work. I want to thank Anindita Sarangi for her continuous support in the lab. I want to thank all the undergraduate students, Dylan Fingerman, Constance Huang, and Hemapriya Dhanasekaran, for assisting me in all my research projects. Lastly, I would like to express my deepest thanks to all the past graduate students of the lab, Dai-Tze Wu, Araba Addaquay, Sriram Aiyer, Leonardo skills they have shared with me as well as the camaraderie inside and outside the lab.

Last but not the least, I would like to thank my friends and family for their relentless support throughout my Ph.D. program. I’m grateful for all the friends I’ve met through the program and school especially, Katelyn Marshall and Valentina Dal Pozzo, and friends I’ve met in the New Brunswick community, Michael Pino, Adalin Ball and Spoorthi Kumar. They have all been instrumental to get me through graduate school. Most importantly, I’m extremely grateful to my parents, brother and sister for all the unwavering support they have showered me all my life especially the last 5 years.

Table of Contents

List of Tables

List of Figures

Introduction

Retroviruses are enveloped RNA viruses that are categorized on the basis of genetic, structural, and pathogenic properties. Its quintessential characteristic is its ability to transcribed double stranded viral DNA from its viral RNA genome by reverse transcription and its subsequent insertion into the host genome. They are considered multipotent biological nanoparticles because of their size that ranges from 100-150 nm in size. Their viral RNA genome ranges from 7-12 kbp in length and are linear, single stranded, nonsegmented and positive polarity. They have at least three major coding regions (*gag, pol, env*) and flanked by long-terminal repeats (LTR) on the 5' and 3'. The LTRs provide the transcriptional control and polyadenlylation elements for the viral transcript as well as the binding sites for IN integration (Fig 1). The *gag* region, in general, codes for at least 3 structural viral proteins: matrix (MA), capsid (CA) and nucleocapsid (NC). The *pol* region codes for catalytic proteins: protease (PR), reverse transcriptase (RT) and integrase (IN). The *env* region contains the viral Envelope protein (1). Retroviruses that carry these three coding regions are called simple retroviruses - i.e. Gammaretrovirus, Murine leukemia virus (MLV) - while retroviruses that carry additional viral proteins formed from alternative splicing are called complex retroviruses - i.e. Lentivirus, human immunodeficiency virus (HIV) (reviewed in (2)).

Retroviruses are further divided into 7 genera - alpharetrovirus, betaretrovirus, deltaretrovirus, epsilonretrovirus, gammaretrovirus, lentivirus and spumavirus- based on evolutionary relatedness primarily on sequence relatedness of the reverse transcriptase (1). Retroviruses can be further classified also based on the morphogenetic assembly of viral particles. Cores of B- and D- type, for example in mouse mammary tumor virus (MMTV) and foamy virus (FV), are assembled in the cytoplasm. C-type viruses like MLV and HIV assemble their core during the budding process (1).

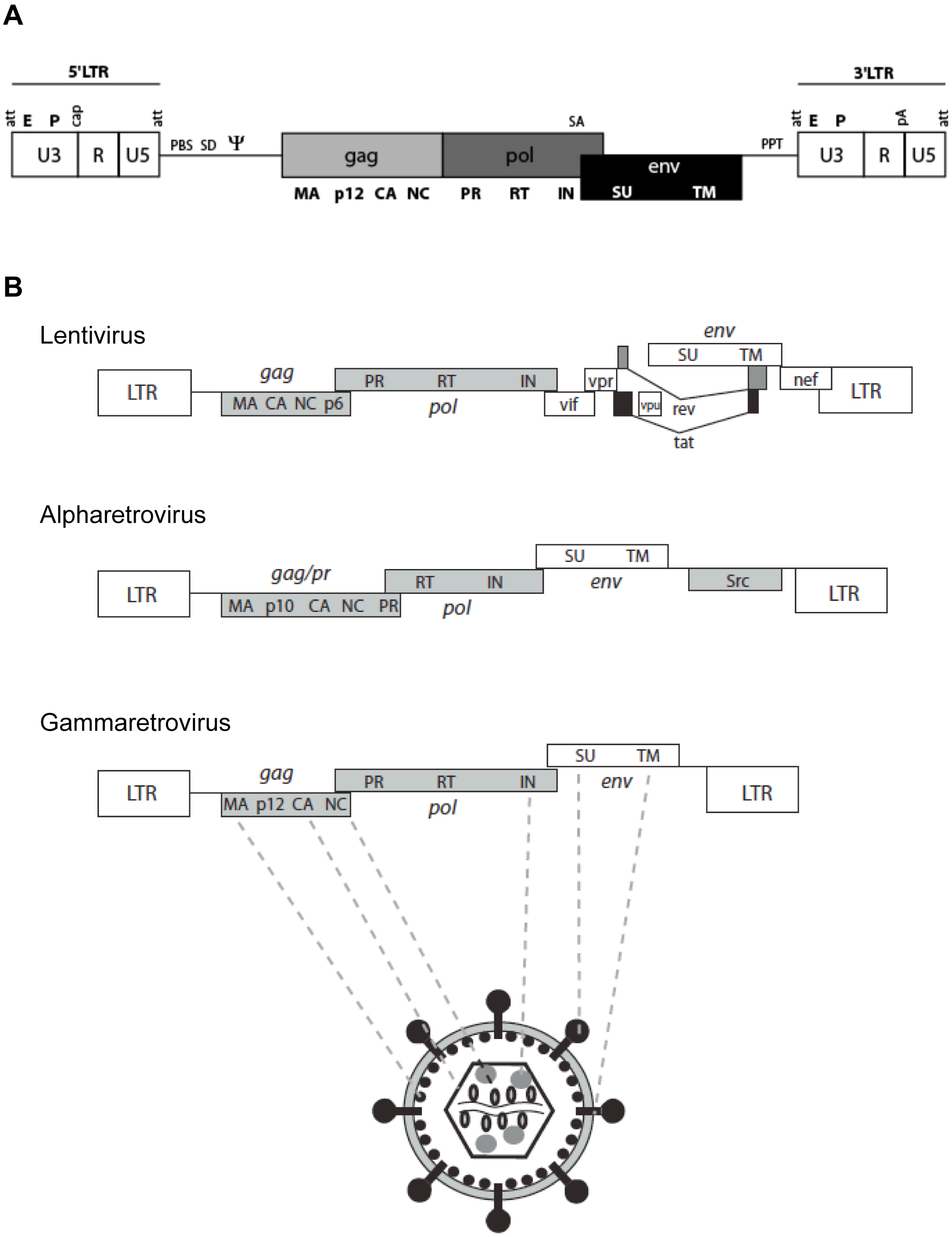


Fig 1. Genome organization of retroviruses. The panels are adapted from (2, 3) (A) Gammaretrovirus (MLV) genome organization is an example of a simple retrovirus. LTR (open boxes) are at the 5' and 3' end of the genome. It consists of the U3, R. U5 regions. Some regulatory elements are also indicated (E, enhancer; P, promoter; att, attachment site; cap, 5'RNA capping site; pA, polyadenylation site; PBS, primer binding site; SD, splice donor; Ψ, packaging signal; SA, splice acceptor; PPT, polypurine tract). Three coding regions are indicated (*gag, pol, env*) and their protein subunits indicated in bold MA, matrix; CA, capsid; NC, nucleocapsid; PR, protease; RT, reverse transcriptase;

IN, integrase; SU, surface; TM, trans-membrane) **(B)** Comparison of different retrovirus genome organizations. Lentivirus encodes accessory proteins (Vif, Vpr, Vpu, Tat, Rev, and Nef) formed by alternative splicing. The alpharetrovirus protease (PR) is encoded as part of *gag* and its genome contain the *Src* oncogene. Typical viral structure is depicted at the bottom.

Retrovirus Genome

Retroviral LTR

The retroviral LTR is composed of different regulatory elements that are essential for viral replication, transcription and integration. It is composed of the U3, R, U5 region (U-unique; R-repeat) and 2 copies of these components flank both ends of the viral DNA genome (Fig 1A). The size varies among different genera. The U3 has the active enhancer/promoter that drives transcription by host polymerase II. This can also define retroviral tropism in which transcription activity is dependent on the infected cell or tissues type. MLV is highly expressed in thymocytes and lymphoid cells (4-8). Present at the 5’ terminus of the U3 is the highly conserved inverted repeat attachment (*att*) site, where 3’ processing occurs at the CA dinucleotide during integration (9). R region is defined from the transcriptional start sites at the 5’ LTR and polyadenylation site at the 3‘ LTR. The 5’ end of the viral RNA transcript begins in R and is capped by a regular m7G5’pppp5”G_mp-cap (10) (Fig 1A). Reverse transcription of the viral RNA produces the complete copy of LTR at each end of the viral DNA before integration. The U5 also encodes an *att* site for IN during integration. Other regulatory elements are present downstream of the 5’ LTR called 5’ untranslated region (UTR). The 18-bp long primer binding site (PBS) serves as the binding site for complimentary host tRNA that serves as primer for reverse transcription (1). This tRNA varies among retroviruses but for MLV, it’s usually $tRNA^{Pro}$ but also can be $tRNA^{Gln}$ (11, 12). Retroviral RNA dimerization signals overlap regulatory elements for RNA packaging. The packaging signal, (Ψ)- site, interacts with nucleocapsid (13). The major splice donor sites (SD) for almost all retroviruses are present upstream or within crucial packaging motifs (1, 14).

Coding Regions

Viral coding regions expressed polyproteins and its transcription leads to different precursors, Gag and Gag-Pol from the full-length RNA genome. These fusion proteins assemble and form immature viral particles and carry the viral RNA genome. At viral maturation, proteolytic processing by protease (PR) recognizes cleavage sites flanking individual viral proteins to allow proper functional configuration (14, 15). Expression of the Gag and Gag-Pol are expressed in 20:1 to 10:1 ratio for efficient production of infectious particles (1). Differential translational expression is achieved by 2 different mechanisms: leaky stops and RNA frameshifting. For gammaretrovirus and episolon retrovirus, the terminal UAG codon at the end of *gag* precedes the *pol* coding region and read-through inserting a glutamine by the suppressor tRNA^{Gln} occurs at a 5-10% frequency. Downstream of the UAG codon, a pseudoknot is formed that is required for efficient read-through (14, 16). The read through is enhanced by reverse transcriptase by binding to eRF-1 (17). For beta-, delta-, lentivirus, a translational -1 frameshift at a 5' direction occurs near the *gag* stop codon and requires secondary structure formation, a hairpin or pseudoknots, to stall ribosomal translation (18-20).

Gag

The Gag polyprotein is composed of structural proteins that are structurally conserved among different genera of the Retroviridae. MLV, in addition to the primary AUG Gag start, encodes the atypical CTG start codon 264 bp upstream of the Gag AUG. This extension codes for a leader peptide that directs a secondary Gag translation (glyco-Gag) to the endoplasmic reticulum and Golgi complex where it is glycosylated, (21, 22). Its presence is non-essential to viral competency and pathogenicity in cell culture. It was reported, however, that absence of glyco-Gag affect infectivity and budding in mice and it directs virion assembly to lipid rafts involving cellular La protein (23-25). Gag mainly consists of at least the MA, CA and NC. The MA is involved with interaction with cytoplasmic tail of Env as it is bound to the viral lipid bilayer following precursor cleavage. CA is a 20-30 kD protein and has the most highly conserved sequences among all *gag* subunits called the major homology region (MHR). The NC is encoded after CA domain and is highly basic. They contain one or more Cys-His motifs that interact with zinc ion and these are essential for viral RNA genome packaging. It's also part of other viral processes in HIV including tRNA binding in assembly, reverse transcription and integration (26, 27). Alpha-, beta- and gamma-retroviruses also code for phosphoprotein (10-20 kD) between matrix and capsid (28). In MLV, this protein is called, p12, and contains the PPPY late domain that interacts with the host E3 ubiquitin ligases, NEDD4 family, to facilitate Gag ubiquitylation and promote viral release (29-32). The C terminal end of p12 binds to CA and is required for tethering of CA containing pre-integration complex (PIC) to host mitotic chromosomes, specifically to nucleosomal histones (29-34).

Pol

The *pol* coding region usually consists of the PR, RT, and IN. PR, a 15 kD homodimer, promotes aspartyl protease activity (35). Some possible mechanisms of activation are decrease in pH environment after viral budding or high concentration for protease dimerization (14). RT is an important enzyme for viral pathogenesis. It has an N-terminal polymerase domain and a C-terminal RNase H domain (36). It is involved in RNA-templated DNA synthesis, DNA-templated DNA synthesis, and degradation of the RNA strand of the RNA:DNA hybrid after DNA is synthesized (14). In MLV, it functions as a monomer unlike other retroviruses (37, 38). It is, however, a polymerase with low fidelity and processivity and does not have a proofreading activity. IN enzyme process 2 catalytic activities: 3' end processing of each DNA strand and strand transfer of viral DNA into host DNA (reviewed in (39, 40)).

Env

The Env is expressed from a spliced transcript for all retroviruses. It is the only viral protein expressed as a surface protein and it functions for viral attachment and recognition. It determines the ability of viruses to infect specific cells or commonly referred to as the viral tropism. The precursor goes through different post-translational modifications, disulfide formation, oligomerization and cleavage into 2 subunits. These subunits are the N-terminal surface subunit (SU) and the C-terminal transmembrane subunit (TM) (41). SU have an N-terminal receptor binding domain or RBD with a proline rich hinge region followed by a conserved C-terminal domain (42). The RBD provides contacts with host surface receptor proteins and therefore contains variable regions. The TM subunit starts with a hydrophobic peptide, called the fusion peptide, which is inserted into the host cell membrane as part of early stage of membrane fusion. It is followed by a stretch of leucine zipper-like regions that form a coil-coil structure (43). MLV Env is tethered together by a disulphide bond by through a cysteine CXXC motif in SU with the CX6XX motif in TM. The RBD during receptor recognition changes its conformation to communicate with the C-terminal domain of SU and TM. This activates the reshuffling of cysteine bond to the cysteine bonds of CXXC. SU is released inducing conformational change in TM and the insertion of fusion peptide into the target membrane (44-46).

Retroviral life cycle: Gammaretrovirus

The retroviral life cycle is divided into 2 stages (Fig 2): early and late stages. The early stage starts with the viral Env binding to receptors on the host cell membrane. This leads to conformational changes that initiate membrane fusion or endocytosis and eventual entry of the viral core into the cytoplasm (reviewed in (2, 3)). Reverse transcription turns the viral RNA genome into a double-stranded DNA molecule that is incorporated into the pre-integration complex (PIC) along with CA, p12, IN and other cellular proteins (barrier to autointegration, Baf protein). The size of MLV PIC, due to the presence CA core, precludes passage through the nuclear pore therefore it depends on the nuclear breakdown during mitosis to gain access to the nucleus (14, 47). Baf proteins prevent intramolecular strand transfer by binding to proviral double stranded DNA and promoting successful integration with host DNA (48). Integration proceeds with a 3-step process that includes host proteins (discussion in the next section): 3' processing, strand transfer and gap repair/DNA strand joining. The integrated viral genome is now called an integrated provirus at this point (reviewed in (39)).

The late stage commences when viral RNA transcription is initiated from the 5' U3 LTR. Transcribed viral RNAs are spliced at the donor sites located at the 5' of *gag* and the splice acceptor site at the 3' end of *pol* (49). A full unspliced RNA can serve as template for Gag and Gag-Pol expression or used for genome replication. Full length genomic RNA is packaged by NC of uncleaved Gag and Gag-Pol precursors. Export elements from the nucleus for MLV full-length genomic RNA transcript has not been identified. Gag and Gag-Pol precursors are assembled into the membrane, specifically at regions called lipid rafts, guided by the myristoylation signal in MA and form protein-protein interactions that guides budding off from the membrane (50-52). Env proteins are also incorporated into the viral membrane shell during budding of viral particles. Particles can be formed without Env and are called virus like particles (53, 54). Maturation of viral particles involves proteolytic cleavage by the viral PR of Gag and Gag-Pol precursors into its individual functional subunits and the replication cycle can begin again.

Retroviral integration

Integration is an essential step for retroviral replication and pathogenesis (for review (39)) This process is driven by the viral integrase (IN), which stably introduces the proviral DNA into the host genome for subsequent viral replication. Other less characterized functions of IN have been observed. Disruption of the coding sequences of IN caused pleiotropic phenotypes including aberrant viral morphology with reduced

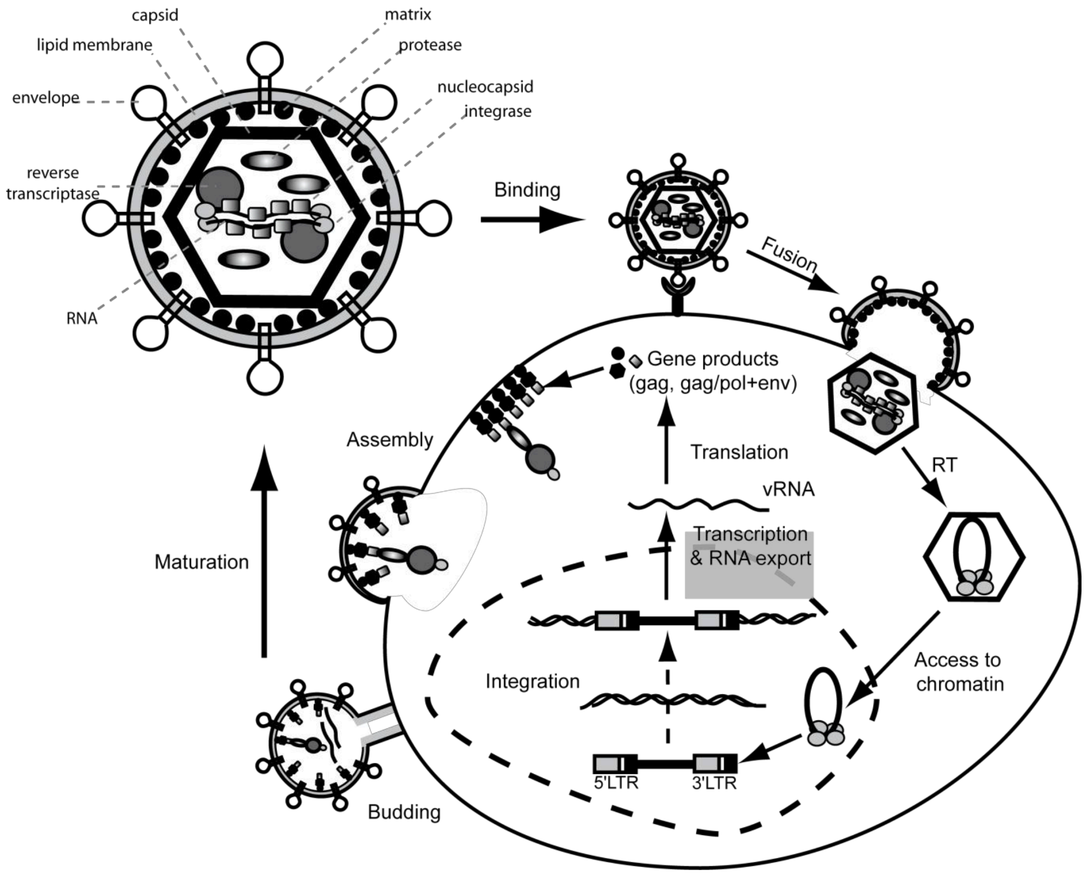


Fig 2. Gammaretroviral life cycle. The figure adapted from (3). The life cycle starts with binding to receptors then entry. Reverse transcription (RT) follows and the viral DNA with RT, integrase, and p12 forms the pre-integration complex (PIC). After nuclear membrane breakdown during mitosis, the PIC can localize into the chromatin and integration can proceed after. Viral transcription and translation produces viral protein precursors and assemble and bud from the plasma membrane. Viral particles mature when the protease cleaves the viral protein precursors into their subunits.

level to no activity of RT, virion assembly disruption and nuclear import defects which suggest it could have other functions outside of integration (55-58)

Integration is an enzymatic process with multiple steps that involves the viral LTR and the host DNA target catalyzed by the viral IN and host enzymes. During infection, some viral DNAs are circularized into 2 LTR-circular forms but the linear viral DNA is the correct substrate for integration. Integration is made up of 3 steps (Fig 3) (reviewed in (39)). Firstly, IN hydrolyzes the phosphodiester bond at the 3' of the conserved adenine of the 5' CAGT- 3' removing a dinucleotide (GT) at both ends of the LTR. This also releases the 3' hydroxyl group from the 5' CA- 3'. In the second step, the intasome, complex of viral DNA and IN, will form the target capture complex (TCC) with target host DNA through a nucleophilic attack of 3' hydoxyl to cut DNA in a staggered manner. A post-catalytic complex called strand transfer complex (STC) is formed where viral DNA is hemi-integrated into the host DNA. Full integration into a provirus occurs when STC is disassembled and completed by host enzymes. DNA polymerase (β or δ) processes the single stranded gaps formed during stand joining. 5'-flap endonucleases 1 excise the dinucleotide 5' overhang in the 5' viral DNA produced after DNA polymerase extension. DNA ligase III, IV, or I then would ligate the joining ends. At the end of integration, provirus is flanked by duplication of the target DNA sequence. MLV and PFV have a 4 bp duplication size while HIV has 5 bp (reviewed in (3, 14)).

MLV IN: structure and function

Integrase (IN) structure is important in understanding its function in viral pathogenesis. MLV IN is 408 residues in size and has 3 domains: N-terminal domain (NTD), catalytic core domain (CCD), and C-terminal domain (CTD) (Fig 4A). The

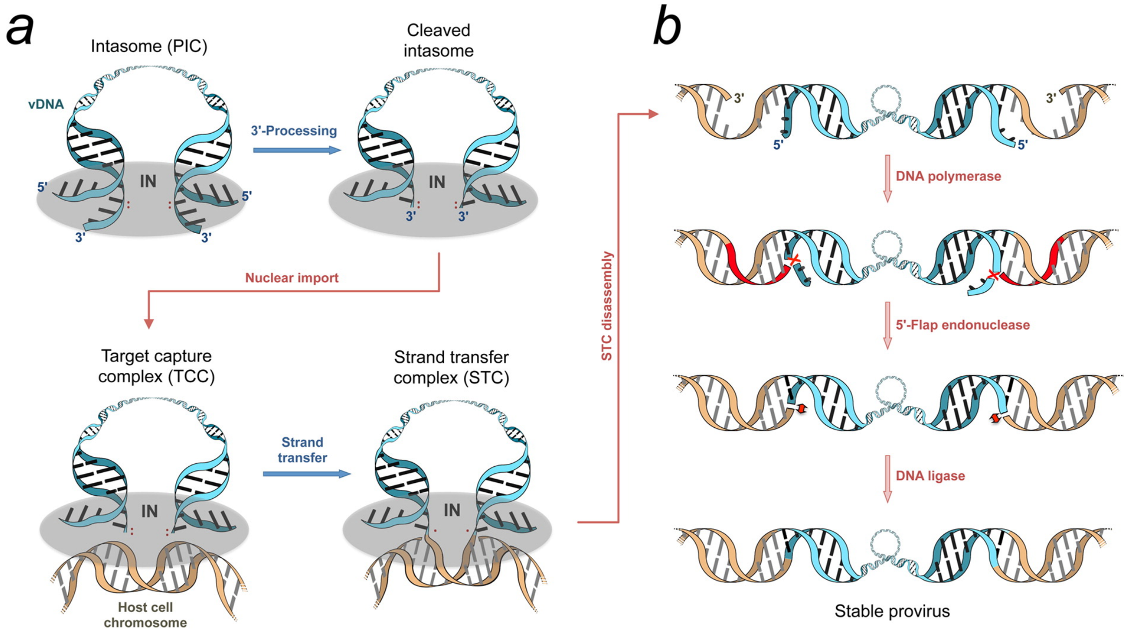


Fig 3. Retroviral integration This figure is adapted from (39). The 3-step integration involves host proteins for DNA repair (B, red arrows). (A) First, the intasome is assembled with the multimer of IN (gray oval) bound to viral DNA ends. 3' processing follows to expose the –OH group then intasome engages with the target to DNA to form the target capture complex (TCC). Second, 3' ends inserts itself into the target DNA by a nucleophilic attack to form the strand transfer complex (STC) with hemi-integrated DNA. (B) DNA repair for strand discontinuities, base excision, and ligation are catalyzed by host proteins.

complete structure of the integrase protein has not been solved, but separately the structures of the N-terminal and C-terminal domains have been determined in our lab using X-ray crystallography and nuclear magnetic resonance (NMR), respectively, and a molecular model for the catalytic domain has been generated by homology modeling (59-61) (Fig 4B). Other retroviruses such as the Rous sarcoma virus (RSV), murine mammary tumor virus (MMTV), and the prototype foamy virus (PFV), have available structures of the their integrase protein (62-65). Cryo-electron microscopy (EM) was used to determine the structure of whole HIV MMTV and the PFV intasome bound to nucleosome (63-65).

MLV and PFV possess an extra domain (45-51 residues) before the NTD, called NED, which engages the viral DNA into the intasome. The NTD forms a helx-turn-helix fold with a histidine and cysteine residues motif (HHCC) bound with a Zn^{2+} ion and is involved in recognition of the viral LTRs and IN multimerization. The NED plus the NTD are collectively called the N-terminal region (NTR). The CCD harbors an RNase H fold and contains the highly conserved active site motif consisting of Asp and Glu (DD-35-E motif). The active site coordinates the Mg^{2+} that activates the nucleophilic attack for 3'-processing and strand transfer to destabilize the phosphodiester bonds (reviewed in (39, 40, 66). There's a flexible linker region between MLV CCD and CTD from residues 287-329 called the CCD-CTD linker (59, 60, 66, 67). The CTD is the least conserved domain and it contains a Src homology 3 (SH3) fold. It is most related structurally with the Tudor family of chromatin binding domains but does not possess the conserved hydrophobic cages that bind the methylated Lys and Arg bonds. The CTD is involved in interaction with DNA and other substrates (14, 59, 60, 68-70). Residues from 382–408 constitute the MLV CTD C-terminal tail peptide (TP) (59, 60, 66, 69). Sequence alignment of CTD TP of gammaretroviral INs showed a conserved sequence (^{390}W-X(3)-R/K-S/T-X(2)-PLK-I/L-R-I/L-X-R^{405}) that's exclusive to gammaretroviruses. The coding

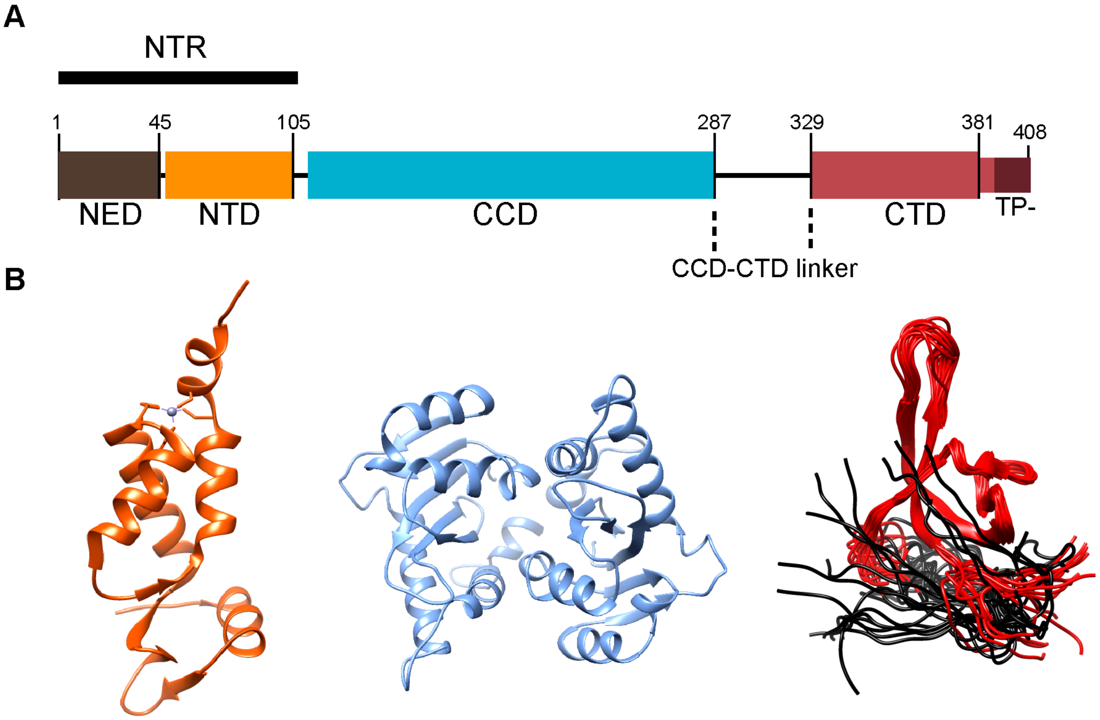


Fig 4. Domain organization of MLV IN. (A) Schematic shows IN protein sequence (brown – NED; orange – NTD; Blue – CCD; red – CTD). TP^- denotes the region deleted in the construct used for mouse infection and integration site analyses. (B) NTR_{1-105} (orange) has subdomains NED and NTD and shows a Zn^{2+} ion (PDB ID: 3NNQ). CCD $_{117-271}$ dimer was based from PFV CCD (59). Overlay of backbone atoms of the NMR structures of CTD domain shows the intrinsic disordered structure, TP (black), of CTD (PDB ID: 2M9U).

region of the N-terminal of Env overlaps with the CTD tail peptide sequence but expressed using a different reading frame. The sequence conservation within the overlap region favors the IN coding sequence. This further proves that there is a selective pressure to preserved the IN C-terminal TP sequence (69). MLV IN is presumed to function as a tetramer as observed in prototype foamy virus (PFV) (71, 72). Structural analysis of multimers of different retroviral IN shows that with spuma-, epsilon- and gammaretroviruses have longer (>50 residues) CCD-CTD linkers that provide sufficient engagement of the CTD to viral and target DNA thus requiring an IN functioning as a tetramer, compared to RSV and MMTV, both octamers (62, 63).

Retroviral integration preferences and host protein interactions

There are retroviruses that have strong preference for specific genome regions. HIV integrations highly prefers active gene bodies (73). Spumaviruses target intergenic regions and avoids gene bodes. Alpa-, beta-, delta- retroviruses have a more random distribution compared to other retroviruses (74). MLV has preferential integration near promoter, enhancer regions and transcription start sites (TSS) that are high transcriptionally active regions (60, 69, 70, 75, 76). These preferences can be explained by cellular host protein that interacts with the viral PIC and guide integration (3, 77-79).

Some of the host protein interactions with some retroviral INs have been identified and structurally and biochemically analyzed. HIV IN tethers to an epithelium-derived growth factor/p75 (LEDGF/p75) that interacts with chromatin specifically H3K36me3, a histone modification mark for transcription elongation (80-83). The IN CTD of alpharetrovirus, avian leukosis virus (ALV), binds to the FACT complex protein, a highly conserved histone chaperone protein essential for transcription and DNA replication, and promotes integration for ALV. The FACT complex has recently been shown to regulate HIV integration (84, 85). The interaction of the MLV IN with the host bromo- and extraterminal (BET) domain proteins influences target-site selection to transcriptionally highly active chromatin regions. MLV IN has an unstructured C-terminal tail peptide (TP) of CTD that becomes ordered upon interaction with BET proteins (60, 69).

BET proteins are known epigenetic readers and are characterized by the presence of two bromodomains (BD1 and BD2), an extraterminal domain (ET), and a C-terminal domain (CTD). Specifically, MLV IN TP interacts with the ET domain. The bromodomain is mainly linked to the histone acetyltransferase (HAT) activity of transcriptional activators to recognize the histone acetylation. BET proteins function as molecular scaffolds to recruit other proteins and are usually localized at promoter and enhancer of active regions and possess nucleosome chaperone activities to promote RNA polymerase elongation (86, 87). Additionally, a BET protein, Brd4, has been localized inside the nucleus during interphase and it's closely associated with chromatin during the M phase of cell division (87).

MLV common insertion sites using next generation sequencing

Integration profiles of retroviruses are important in identifying potential insertional mutagenesis targets for cancer gene discovery. Common insertion sites (CIS) of MLV were first identified by comparing the integrations versus the randomly generated integrations from 100,000 Monte Carlo trials between B-cell lymphoma and myeloid leukemia (88, 89). Other CIS analysis enhanced identification of insertions at genes to further predict oncogenic potential (90). The most recent CIS analysis compared previous CIS data sets of different genetic backgrounds and their data set and described a progression network of MLV target genes that has strong indication for oncogenic potential beyond the known MLV integration preference (90).

Insertional mutagenesis of retroviruses

In mice, Moloney MLV (M-MLV) is non-acute retroviruses and thus insertional activation of proto-oncogenes at identified common insertions sites (CIS) is the predominant mechanism of oncogenesis (91, 92), requiring a long-latency period varying between 4-12 months (91). There are many insertion mechanisms that retroviruses use to drive oncogenesis (Fig 5). First is through promoter insertion, where the integrated provirus is in the same orientation of the gene at the promoter region at the 5' end and activating the gene through the viral LTR instead (Fig 5A). Second is through an intragenic insertion, where it could either produce a truncated transcript via the poly-A tail signal or a chimeric transcript if integrated within the exon (Fig 5C). Disrupting transcripts can lead to inactivation of a gene, for example insertion within a tumor suppressor gene (e.g. *p53*) can lead to loss-of-function (93-95). The chimeric protein could be a gain-of-function and cause aberrant effects on the cellular activity. Lentiviral vectors mostly demonstrate this mutagenesis effect for example; it caused an activated form of a truncated HMGA2 proto-oncogene in hematopoietic cells in patients being treated with beta-thalassemia (96). The last mechanism is enhancer insertion where integration occurs at distance upstream from a gene and usually oriented at anti-sense of the gene but it could also be in the sense. MLV can influence expression of genes that are >100kb away from integrants (97, 98) (Fig 5B).

MYC/Rux2 tumorigenesis mice model

A transgenic mouse overexpressing two MLV CIS, *MYC* and *Runx2* from CD2 promoters leads to early onset lymphomagenesis. The synergistic expression of *MYC* and *Runx2* is proposed to neutralize p53 activation by modifying each other's effect on the p53 pathway. Varied expression of human-CD2-*MYC* in the CD-*MYC* mice model can lead to MYC-induced apoptosis. Its expression influences the T-cell repertoire and

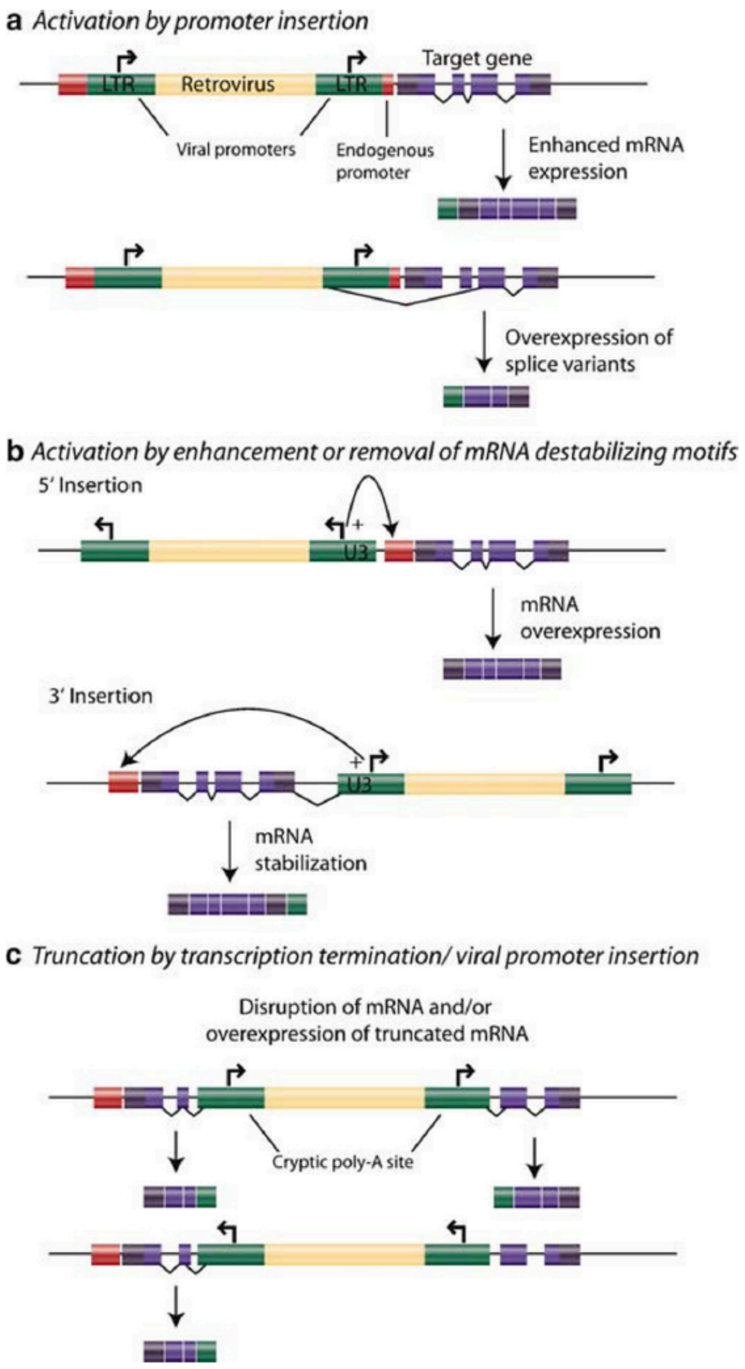

a Activation by promoter insertion
LTR
Retrovirus
LTR
Target gene
Viral promoters
Endogenous promoter
Enhanced mRNA expression
Overexpression of splice variants
b Activation by enhancement or removal of mRNA destabilizing motifs
5′ Insertion
+
U3
mRNA overexpression
3′ Insertion
+
U3
mRNA stabilization
c Truncation by transcription termination/ viral promoter insertion
Disruption of mRNA and/or overexpression of truncated mRNA
Cryptic poly-A site

Fig 5. Mechanisms of insertional mutagenesis. This figure is adapted from (95). Provirus is yellow while LTR is green. The blue boxes are exons of genes and red boxes are the endogenous promoter. (A) Promoter insertions (B) Enhancer insertions in sense and anti-sense directions (C) Intragenic insertions in exons and introns.

proliferation by enhancing positive selection and expression of T-cell receptors. Co-expression with the CD2-*Runx2* at the same stage of lymphoma development reduced the rate of apoptosis for the transgenic mice. In the CD-*Runx2* mice model, it showed growth suppressive effect in thymic cells development. Co-expression with CD2-*MYC*, counteracts this suppressive nature and increased the onset of tumors (99, 100). Infection of WT MLV into *MYC/Runx2* mice reduced survival by 10 days (90, 100). Additional neonatal infection of this mice model with M-MLV WT virus accelerates tumorigenesis and increases clonal complexity through various insertional mutagenesis sites (90). Analysis of these integration sites through next-generation sequencing, mapping against reference genomes and ChIP-seq data sets, identified a panel of MLV CIS that accelerated the oncogenic process (90).

Recombination with endogenous viruses

Murine gammaretroviruses are classified based on the exogenous versus endogenous localization and receptor usage (101, 102). Inbred strains of mice encode endogenous type C murine leukemia viruses. These fall in three general classes, differing in their receptor usage and thus their host and tissue specificities. These include the ecotropic viruses, limited to rodents (mCAT1 receptor), xenotropic viruses (excluded from infection of mice; Xpr1 receptor), and polytropic/modified polytropic (103) with a broad host range (101, 102) and the Xpr1 phosphate exporter as receptor (104). Although endogenous ecotropic and xenotropic viruses can form infectious particles, the polytropic MLVs (P-MLVs) do not produce replication competent viruses (101) due to defects in the ORFs, insertions, and mutations within the LTRs.

These endogenous sequences, however, are an abundant source for recombination, when challenged with alternative defective viruses or replication competent viruses. Endogenous retroviral RNA can be co-packaged with exogenous retroviruses where high rate of homologous recombination can occur during reverse transcription (Fig 6). If there are breaks in the RNA genome during minus strand DNA synthesis, the co-packaged RNA genome can be used as the new template (forced copy-choice mode) (Fig 6B). Another mechanism is when an internally initiated plus strand DNA fragment jumps to another minus strand DNA molecule and assimilated into the viral DNA (Strand-displacement-assimilation model) (Fig 6A) (105-108). The most favored mechanism involves base pairing of nascent DNA, exposed after RNase H degradation of a DNA:RNA template, with copackaged genomic RNA. Primer strand realignment then DNA synthesis proceeds into an acceptor template that will produce a recombination junction (Minus-strand exchange model) (Fig 6C) (109). In MLV, polytropic and modified polytropic ERVs can contribute *env* sequences in recombination with breakpoints/crossover located in IN and around TM region of Env (103, 110). The generation of these recombinants frequently results in viruses with improved virulence and exchange of the viral *env* (7, 111, 112). Of significance to this study, C57BL/10 express the xenotropic MLV from the *Bxv-1* locus, which can be a source of viral proteins as well as genetic material (113).

Gene therapy

Historically, MLV-based vectors were used in the initial human gene therapy trials (114). Gammaretroviruses have a simple genome organization, high transduction rates, broad tropism and can only infect dividing cells which make them suitable as gene therapy vector. In multiple clinical trials, including X-linked SCID (115-117), X-linked chronic granulomatous disease (118) and Wiskott-Aldrich Syndrome (119), but not ADA deficiencies (114), insertional mutagenesis resulted in the outgrowth of oligoclonal populations due to trans-activation of proto-oncogenes (120).

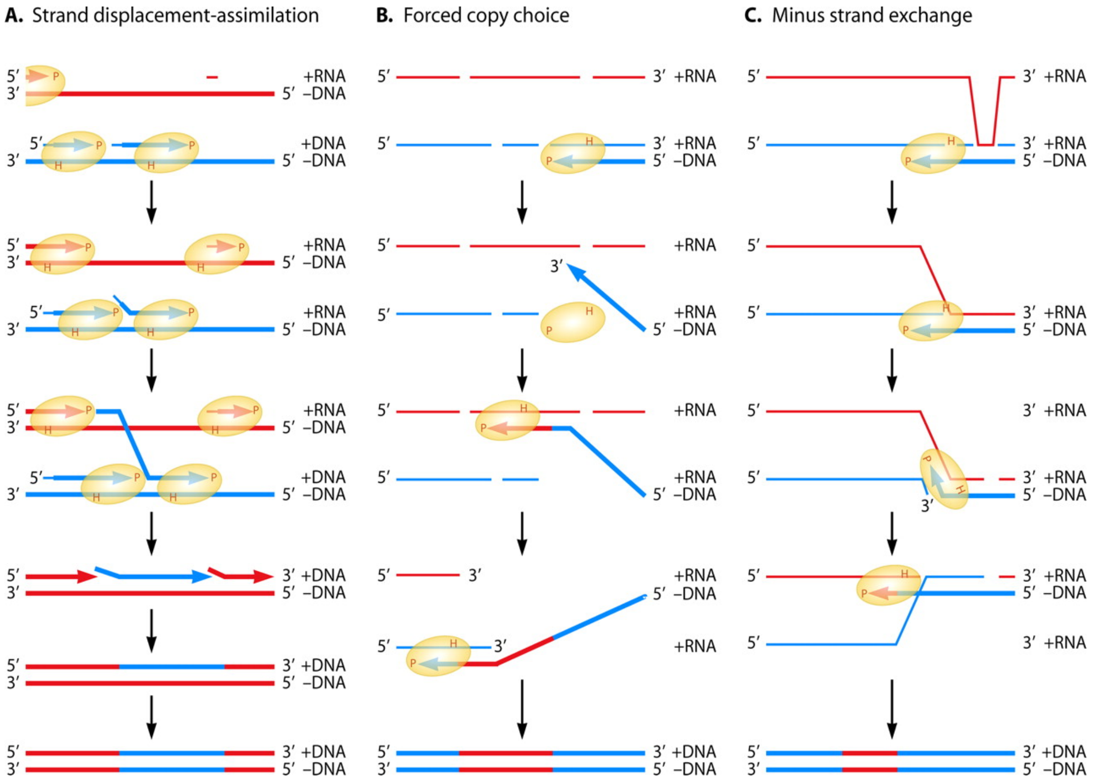


Fig 6. Three models of recombination. This figure is adapted from (109). (A) Strand displacement-assimilation occurs during plus strand synthesis. It is initiated in multiple positions and strand displacement produce free DNA tails that could move into a different strand. (B) Forced-copy choice model shows nascent DNAs encounter nicked ends; primer strands can associate with homologous regions the co-packaged genomic RNA. (C) In minus-strand exchange model, nascent minus-strand DNA is exposed after RNase H template degradation and it is free to pair to complementary region of the other genomic RNA then DNA synthesis follows.

These clinical trials showed that integration of MLV vectors can lead to premalignant clonal expansion that subsequently increased the number of mutations. There are common insertion sites such as *LMO2* locus, a proto-oncogene that caused childhood T-cell leukemia. For some patients in SCID-X1 and WAS trials, T-cell malignancies were driven by integrations at this locus (121). If the all the SCID-X1, ADA-deficient SCID and WAS clinical trials are considered, insertions at this locus cannot be the only pathway towards transformation of T-cells and driving progression of malignancy (116, 122, 123). There are many insertional mutations found in patients that increased the complexity in the development of oncogenesis. Many other factors during treatment can also have influenced progression of oncogenesis in individual patients such as disease context, patient's genetic background or the MLV vector design (reviewed in (94))

Genotoxicity of MLV vectors as demonstrated from these past clinical trials have led to studies to decrease its oncogenic potential. Many alternative approaches have been developed including using self-inactivating vectors (124) or lentiviral vectors (114). Another approach is inserting peptides or protein domains that interact to specific region of the genome to redirect integration (125, 126). Alternatively, addressing the integration target-site bias of gammaretroviruses to integrate preferentially at promoter/enhancer regions could alter the oncogenic potential of these vectors (60, 125)

Objectives/Rational

Disrupting the host protein interactions with retroviral IN decreases the bias of integrations to regions with high oncogenic potential. In the case of MLV, removing the C-terminal TP that interacts with the ET domain of BET protein decreased the integrations significantly at high transcriptionally active regions. These studies were performed in cell culture to demonstrate the influence of BET proteins with MLV integration preference. Engineering MLV IN to change integration preference, therefore, is a strategy to improve MLV as gene therapy vectors. This observation must also be validated in animal model to consider effects of endogenous elements on retroviral pathogenesis and address the selective pressure for tumor outgrowth.

This is the first study to demonstrate infection of the replication-competent MLV without IN TP (IN TP$^-$) in a transgenic mouse tumorigenesis model (*MYC/Runx2*) and examines the effects on tumorigenesis and the MLV integration profile. The study highlights the strong selective pressure on MLV to maintain the IN TP, through either internal deletions or recombination with endogenous retroviruses. MLV IN TP$^-$ integrations in tumors from *MYC/Runx2* mice were mapped using next-generation sequencing (NGS) using ligation mediated PCR and correlated to known binding sites of BET protein and histone modification ChIP-seq peaks. Additionally, tumor progression and insertional mutagenesis in *MYC/Runx2* mice infected with MLV maintaining the IN TP$^-$ genotype was analyzed. Furthermore, MLV IN TP$^-$ integrations in human K562 cells were characterized with respect to different chromatin states and different histone modification ChIP-seq peaks.

MATERIALS AND METHODS

Cells lines.

293T cells, 293mCAT cells (expressing mCAT receptor) (30) and D17/pJET cells (expresses the mCAT-1 receptor) (127) were maintained in Dulbecco's modified Eagle medium (DMEM; Gibco #11965) supplemented with 10% (vol/vol) heat-inactivated fetal bovine serum (Atlanta Biologicals # S1245OH) and 1x × antibiotic-antimycotic (100 units/mL of penicillin, 100 µg/mL of streptomycin, and 0.25 µg/mL Amphotericin B) (Gibco #15240). K562 cells were acquired from ATCC (CCL-243) and maintained in Iscove's modified Dulbecco's medium (IMDM, # 12440079) supplemented with 10% (vol/vol) heat-inactivated fetal bovine serum (Atlanta Biologicals # S1245OH).

Plasmids and vectors construction

The replication-competent M-MLV proviral construct pNCA-C (128) and pNCA-C IN-XN (previously named *in*6215a (129)), bearing a 23-aa truncation of the IN tail peptide (TP) of the C-terminal domain (CTD) was previously described (129). To generate a codon-optimized pNCA-C-TP^{-}, a 137 bp gene block (IDT) was chemically synthesized and amplified using primers NCACXN_ScaI6330_rev and NCACXN_NotI6220_fwd. Overlapping PCR of this fragment with a ScaI-ClaI fragment from pNCA-C (generated using primers NCAC_8290_rev and NCACXN_6327_fwd) resulted in a NotI-ClaI fragment, which was exchanged into NotI/ClaI digested pNCA-C IN-XN. Generation of the pNCA-C IN D184N was previously described (130). Sequences of all oligonucleotides are provided in Table 1.

DEAE-dextran transient transfection of proviral DNA clones in D17pJET cells.

Transient expression of the pNCA-C based proviral constructs was performed as previously described (59, 131) using 500 ng pNCA-C based plasmids. Tissue culture supernatant was monitored for viral spread using enzyme-linked immunosorbent assay (ELISA) against MLV p30 (132). Cultures were maintained for at least 14 days prior to analysis.

LacZ viral titer assay

2 x 10^6 293T cells were transfected with 0.8 µg pMD2.G (Addgene) expressing the vesicular stomatitis virus glycoprotein (VSV-G), 0.8 µg pRT43.2Tnlsβ-gal (133) a retroviral packaging vector expressing *lacZ*, and 0.8 µg M-MLV viral genome with wild-type (pNCA-C) or MLV TP^- (pNCA-C TP^-) using Fugene 6 (Promega #E2691) overnight as directed by the manufacturer (134). Viral supernatant was collected, filtered through a 0.45 µm syringe and viral particles were quantified using ELISA against the MLV p30 (132). $1x10^5$ cells D17 cells on 3.5-cm gridded plates were infected with media containing 10 ng of CA in 2 mL DMEM in the presence of 8 µg/ml polybrene. Medium was replaced with fresh DMEM after 24 hrs of infection. Cells were stained for LacZ expression as previously described (135).

Western Blot

Viruses were collected from D17/pJET viral producer cell lines. For CA and Env western blot, 2 mL of viral supernatant was spun at 15,000 x g for 30 min and the viral pellet was resuspended in 20 µL phosphate buffer saline (PBS). Samples were run on a 10 % SDS-PAG and transferred to polyvinyl difluoride (PVDF) membranes using Bio-Rad Trans-Blot® Turbo™ Transfer System. Immunoblots were developed using goat anti-p30 (CA) (1:2000, 81S-263) and goat anti-Env (1:1000, 80S-019) (Quality Biotech) with bovine anti-goat HRP (80S-035-180) as secondary antibody (1:10,000). For IN western blot, 10 mL of viral supernatant was pelleted at 15,000 x g for 30 min and the proteins were visualized using 1:1 mix (1:1000) of antiserum from Rabbit 3 and Rabbit 4, Bleed 5 with goat anti rabbit HRP (Pierce #31460) as secondary antibody (1: 5,000) (136).

MLV IN CCD Purification

MLV IN CCD (residues 106-328) was amplified from plasmid pNCA-C by overlapping PCR to include the C209S mutation (MLV IN $CCD_{106\text{-}328}$ C209S). It was cloned into a pET-15 NESG protein expression vector. Purification was performed based on previous described methods with certain steps optimized for this protein (137). Bacteria with the expression vector were grown overnight in LB with carbenicillin and an aliquot was inoculated in new LB media to make 1:50 dilution in 500 mL. The OD_{600} was monitored until absorbance is between 0.6-0.8 then protein expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at 17°C for 25 hr. Cells were collected and pelleted by spinning at 6000 rpm for 20 mins. Cells were lysed by solubilization buffer (50 mM sodium phosphate pH 7.5, 1 M NaCl, 10 mM 3-[(3-Cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS), 1 tablet protease inhibitor, 20 μM imidazole) and homogenizing the cells with a Dounce homogenizer. Soluble and insoluble fractions are separated by centrifugation for 12,000 rpm for 70 minutes. Soluble fraction was bound to pre-equilibrated (in solubulization buffer for 1 hr; 4 mL of slurry for 500 mL bacterial culture prep) Ni-NTA resin purification (Qiagen # 30250) as suggested by the manufacturer for 2-4 hrs. The protein:resin solution was transferred into a purification column and unbound protein flowed through the column. The column was washed with 20 column volume (CV) of wash buffer (50 mM sodium phosphate pH 7.5, 300 mM NaCl, 10mM CHAPS, 10% glycerol, 50 mM imidazole. 2-step elution was performed with 2 different concentration of imidazole starting with 200 mM followed by

400 mM in 10 mL elution buffer (50 mM sodium phosphate pH 7.5, 1 M NaCl, 10% glycerol, 1 mM 2-mercaptoethanol, immidazole). Zwittergent 3-12 was added into the eluted protein to a final concentration of 0.5%. Buffer exchange followed to remove imidazaole and transfer to the cation exchange chromatography buffer (25 mM sodium phosphate pH 7.5, 100 mM NaCl, 0.5% Zwittergent 3-12). Proteins were loaded into the HiTrap™ SP HP 1 mL (GE) in AKTA FPLC and eluted with gradient concentration (100 mM to 2 M NaCl) of cation exchange chromatography elution buffer (25 mM sodium phosphate pH 7.5, 2M NaCl, 0.5% Zwittergent 3-12). Eluted fractions were ran in SDS-PAG to identify and pool fractions with high amount and purity of protein. Pooled fractions were concentrated using the Amicon® Ultra-4 Centrifugal Filter Unit-10KDa cutoff and Amicon® Ultra-2 Centrifugal Filter Unit-10KDa.

Hydrogen-Deuterium Exchange Mass Spectrometry (HDX-MS)

MLV IN $CCD_{106-328}$ C209S was prepared to 1 mg/mL in 50 µL and detergent reduced by buffer exchange. Sample was submitted to the Rutgers Biological Mass Spectrometry Faculty for HDX analysis. Previously described method was performed on the protein sample (138-140). Incubation period with the buffer in 99.96% ^2H_2O were 10, 100, and 1000 seconds and quenched in 30 µl of 2 M urea, 0.8% formic acid and 50 mM Tris (2-carboxyethyl) phosphine (TCEP).

Infection of MLV into *MYC/Runx2* mice

The *MYC/Runx2* transgenic mice are on a C57/BL6 x CBA/Ca strain background. Infection and maintenance of the mice was as described (141). For WT and mutant MLV, viruses were obtained from 293mCAT cells to avoid recombination with endogenous viruses prior to infection. Briefly, virus isolated from tissue culture supernatant (10^5 TCID50) was inoculated intraperitoneally into mice within 24 hours of birth. Date of Death (DoD) was monitored over a 115-day period (Appendix 1 and 2). For each mouse, a tissue fragment was thawed from liquid N2, chopped up and incubated in medium at 37^{o}C for 2-3 hrs. The medium was spun at 1,200 rpm and then filtered (0.45μm) before adding to 293mCAT cells. Cells were cultured for at least 7 days before harvest for DNA isolation. Amplification of the integrated MLV genomes from the 293mCAT cells was performed using primers 4924 and 7791, previously named 3807 and 6320, respectively (135) (Table). PCR products were cloned using TA cloning. Individual colonies from mice XN-2, 3, and 35 were selected and sequenced for presence of the TP coding region.

Detection of mouse DNA.

Genomic DNA isolates from 293mCAT cells were analyzed for mouse DNA contamination by examining for mouse intracisternal particle A (IAP) and mouse mitochondrial cyclooxygenase-2 (mCOX2). The primers used were mouse_IAP_fwd (ATAATCTGCGCATGAGCCAAGG) and Mouse_IAP_rev (AGGAAGAACACCACAGACCAG) (142), and primers used for COX2 were Mouse_mt__COX2_fwd (5′ TTC TAC CAG CTG TAA TCC TTA 3′) and Mouse_mt_COX2__rev (5′ GTT TTA GGT CGT TTG TTG GGA T 3′) (143). PCR analysis was performed using KOD HotStart.

Single round of infection of K562 cells

Transfection of 293Lenti-X cells using Mirus *Trans*IT®-Lenti transfection reagent with WT MLV and MLV TP$^-$ along with pMD2.G (Addgene) generated viruses for single round infection (144). Viruses were quantitated using ELISA as previously described (132). K562 cells (5x10^5 cells) were prepared a day prior for infection in 6-well plates. For WT MLV, 500 ng of p30 measured was added to one well of K562 cells and for MLV IN

TP^{-}, 5000 ng was added. The plate was spinoculated at 1,500 g for 1 hr then incubated for 4 hrs at 37^{o}C (145). Supernatant was removed and the cells were grown for 24 hrs. The cells were collected for genomic DNA extraction (see below).

Integration sites quantification

Quantification of integrants within in K562 in genomic DNA was performed by quantitative PCR using primers in LTR (5'-AAGAACAGATGGAACAGCTGAATATG-3' and 5'-GCGAACTGATTGGTTAGTTCAAATAA-3') and normalized to the human RPPH1 gene (5'- CGTGAGTCTGTTCCAAGCTC-3' and 5'-GGGAGGTGAGTTCCCAGAG-3') to primers to verify presence of viral integrants. DNA sample (100 ng) were prepared with PowerUp™ SYBR® Green Master Mix in MicroAmp™ Optical 96-Well Reaction Plate as recommended by the manufacturer. Samples were loaded and analyzed using the QuantStudio™ 3 Real-Time PCR System.

Next generation library (NGS) preparation

Genomic DNA from infected mouse tumor, thymus, 293mCAT, and K562 cells were extracted (QIAGEN #6941) and used to generate libraries for MLV integration sites. The protocol for library preparation was adapted from (144). Genomic DNA sample (5 μg) were subjected to two rounds of sonication with the following parameters: duty cycle: 5%; intensity: 3; cycles per burst: 200; time: 80 sec). Purification of DNA for the next generation library protocol used MinElute Reaction Cleanup kit (Qiagen 28204). The sonicated DNA was purified and ends of DNA fragments were repaired using End-It™ DNA End-Repair Kit (ER0720) as described in manufacturer's protocol and purified after. epaired DNA ends were A-tailed using Klenow Fragment (M0212S) and purified. All kits were used as described by the manufacturer's protocol. Illumina linker top and bottom strands (Table 1) were annealed by heating to 90 °C and slowly cooling to room temperature in steps of 1 °C per min. The annealed linkers were ligated with assigned genomic DNA sample with 3000 U of T4 ligase (M0202M) overnight at 12 °C and purified. The first round of PCR used a LTR specific primer and linker specific primer (Table 1) with adapter sequence and primer binding sequence adapted from (144). For the mouse DNA samples, the MLV_LTR_U3_rev primer (5'-GCGTTACTTAAGCTAGCTTGCCAAACCTAC-3) was used (145). For 293mCAT and K562 cells, MLV_LTR_U5 primer (5'-CCTTGGGAGGGTCTCCTCTGAGT-3') was used. Four PCR reactions of 100 ng DNA each were setup for each genomic DNA sample with PCR KOD Hotstart polymerase (Millipore, 71086) under these parameters: One cycle: 98 °C for 2 min; 30 cycles: 98 °C for 15 sec, 60 °C for 30 sec, 70 °C for 45 sec. The reactions were pooled and purified. The second round of PCR used a second round LTR specific primer and the same linker specific primers. These second round LTR specific primers encode for a 6-nucleotide index or barcode sequence compatible for NGS and also encode adapter sequence for DNA clustering and sequencing primer binding site (Table 1). Same number of reactions in the first round PCR was set up for the second round of PCR with the same PCR parameters. All reactions were pooled and purified. Libraries were sequenced and analyzed at the Molecular Biology Core Facilities at the Dana-Farber Cancer Institute.

Library quantification

The NGS library were quantified before Illimuna sequencing. These libraries would contain flanking P5 and P7 flow cell adapters used for sequencing. KAPA library quantification kit contains KAPA SYBR FAST qPCR Master Mix (2X), Primer Premix (10X) (Primer 1 – 5'-AATGATACGGCGACCACCGA-3' and Primer 2 – 5'-CAAGCAGAAGACGGCATACGA- 3') and a set of 6 DNA Standards. Samples were prepared and data analyzed as recommended in manufacturer's protocol. AB StepOne plus instrument was used and ROX high concentration was included in the reaction as an internal control. Library concentration was calculated and melting curve analysis verified size of library fragments and helps differentiate the library from adapter-dimer and other contaminants. Libraries were sequenced and analyzed at the Molecular Biology Core Facilities of Dana-Farber Cancer Institute.

Bioinformatic analysis to identify integration sites from NGS library

Bioinformatic analyses of integration sites were performed as described in (146). LTR and linker sequences are cropped from paired end reads using custom Python scripts, and the cropped reads to the reference genome (mm10 for mouse samples and hg19 for libraries from human cell lines) using HISAT2 (147). Results were then filtered to retain high-quality alignments using SAMtools (148) and unique (deduplicate) integration sites were extracted and formatted to the browser extensible data (BED) format using custom Python scripts. Top-targeted MLV integration sites from the tumor samples was calculated by analyzing the copy number of individual integrants post filtering of high-quality alignments and prior to deduplication using custom R scripts.

Correlation of integration sites with TSS, CpG islands, Brd4, and histone modifications

Integration sites obtained from the tumor of uninfected mouse (NC) was considered as background amplification of the endogenous retroviruses obtained using this pipeline, and hence integration sites from all tumor samples overlapping with the sites from NC were computationally removed. BEDtools software suite (149) was then used to correlate unique integration sites proximal to genomic annotations such as TSS and CpG islands obtained from the University of California Santa Cruz database (http://genome.ucsc.edu/cgi-bin/hgTables). Fraction of integration sites enriched at chromatin associated with Brd4 binding sites and various histone modifications (Appendix 3) was also computed using BEDtools suite. Distance of integration sites from TSS, CpG islands, and Brd4 binding sites were calculated using BEDtools and histograms comparing the obtained distribution of integrants were plotted using ggplot2 (150).

Analysis of integrations with respect to 15- chromatin segmentation states in K562 cells

Genomic annotations showing the chromatin state segmentation of K562 (wgEncodeEH000790) defined by HMM from ENCODE/BROAD was downloaded from the UCSC genome browser. Custom R scripts were written to segregate the individual chromosome state definitions from the master file and BEDtools was used to correlate integration sites within each chromosome states.

Analysis of retroviral integrations overlapping with H3K27ac peaks in mice tumor

H3K27ac ChIP-seq narrow peak files from C57BL/6 mouse thymus were obtained from ENCODE (project code ENCFF001KYC). Overlap between the H3K27ac peaks and retroviral insertion sites for each sample were assessed using the R Bioconductor packages ChIPpeakAnno and GenomicRanges. Distances between H3K27ac peaks and retroviral insertion sites were mapped using the distanceToNearest() function from the R Bioconductor package GRanges, then density plots were produced using the ggpubr and ggplot2 packages. The statistical significance of differences between distances was assessed using Wilcoxon Rank Sum tests.

Chromatin profiling of integrations in K562 and enrichment analysis with histone modification ChIP-seq data

15-state chromatin profiling of integrations of WT MLV and MLV IN TP^- in K562 cells was analyzed (151). All available histone modification data for K562 cells (H3K27ac, H3K9me3, H3K27me3 and H3K36me3, H3K4me3, H2A.Z, H3K9Ac, H3K9me1, H3K4me1, H3K79me2, H3K4me2, H3K4me3, Brd4, CTCF) were used for enrichment of integrations, which is calculated by dividing the number of integrations within a histone modification peaks (+/- 1 kb) by the number of RIC integrations in that same histone modification.

Analysis of M-MLV recombination with endogenous retroviruses

Detection of recombination with ERVs utilized primers recognizing M-MLV, polytropic, xenotropic ERV and the RT_universal_primer (Supplement Table 1) (152). PCR analysis on genomic DNA of 293mCAT virus infected cells was performed using a combination of RT universal primer (5' CCTACTCCGAAGACCCCTCGA-3') and primers specific for polytropic and xenotropic ERVs (Supplement Table 1) using KOD Hotstart polymerase (Millipore, 71086) according to suggested parameters. PCR products from the reaction with RT_Universal_primer and Polytropic_JS5_rev on 293mCAT infected cells from TP^- 6, 7 and 9 mice were cloned into pCR4-TOPO vector from TOPO TA kit following the protocol provided by the manufacturer (Invitrogen, K4575-40). Recombinant plasmids produced were sequenced using the T3/T7 sequencing primers from the manufacturer, and 4981_fwd and MLV_IN_159A_fwd to determine the 5' recombination junction. PCR with Polytropic_JS5_fwd and 7791_reverse primers determined the 3' recombination junction for the same 293mCAT samples. PCR analysis on genomic DNA from mice tumor or thymus samples required a nested PCR. First round PCR used the

RT_universal_fwd primer and the MLV_LTR_U3_rev primer. Second round of PCR used a primer pair of 7791_reverse primer and the Polytropic_JS5_ fwd primer.

Amplification of MLV integrants at targeted loci

TP^{-}16 integration-specific primers were designed based on the genomic location of integrants mapped using next generation sequencing (Supplement table 1). First round PCR used forward primer at the mouse genomic sequence upstream to the integrant and MLV_LTR_U3 reverse primer using PrimeStar GXL DNA polymerase (Takara R050A) or RT_universal_fwd primer and the TP^{-}16 specific reverse primer at the mouse genomic sequence downstream to the integrant using KOD HotStart polymerase. Second round PCR used the same TP^{-}16 specific mouse genomic primer and either primers MLV 20R_reverse or NCAC 6327_reverse or NCAC 5166_forward. For some integrants (*Hdac6* intron 28 and near *Ccnd1)*, additional single linear amplification PCR to amplify the first round PCR was included (153).

Mutagenesis of CCD

The three residues (MLV IN E266, L268, Y269) implicated for BET protein binding were substituted to alanine using overlapping PCR with KOD polymerase (154). PCR of the first fragment was amplified with primer 102510NdeIINteinIN forward and point mutant specific reverse primer (E266A_rev, L268A_rev, Y269A_rev) and the second fragment was amplified using the point mutant specific forward primer (E266A_fwd, L268A_fwd, Y269A_fwd) and the 102510XhoIInteinIN1-407 reverse. The overlapping PCR fragment was introduced into pNCA-C using the HindIII and PmlI sites.

Table 1. List of oligonucleotide primers

Primer	Primer Sequence (5' ->3')	Reference
Virus class specific primers		
Polytropic_JS4_rev	GCAGCCTCTATACAACCTGGGACGGGAG	(152)
Polytropic_JS5_rev	GCAGCCTCTATACTCCCTGAGACTGCCC	(152)
Polytropic_JS5_fwd	GCAGCCTCTATACTCCCTGAGACTGCCC	
Polytropic_JS6_rev	ACGGTCTCTATGGTACCTGGGGCTCCCC	(152)
Polytropic_IN_fwd	TAAAAGCGGCGACAACCCCTCC	
Xenotropic_JS10_rev	ACGGTCTCTATGGTGCCTGGGGCTCCCC	(152)
Amphotropic _rev	ATTCATGGCTCGTACTCTATGGGTTTTAGC	
RT_universal_ fwd	CCTACTCCGAAGACCCCTCGA	
Moloney MLV primers		
MLV_LTR_U3_rev	GCGTTACTTAAGCTAGCTTGCCAAACCTAC	
MLV_LTR_U5_rev	CCTTGGGAGGGTCTCCTCTGAGT	
MLV_IN_159A_fwd		
4981_fwd	GGCTAGAGGCAACCGGATGG	
7791_rev	ccttaaggCCCCCCTTTTTCTGGAGACTAAATA[1]	(135)
4924_fwd	gatatacatatgGCCGTTAAACAGGGA[2]	(135)
6319_rev	AGTACTGCTTCGCCCGGCTCCAGTCCTCA	
TP$^-$16 integrants-specific primers		
HDAC6_ex3_fwd	TCCAGTACCAACTGGCACTCTTGT	
HDAC6_ex3_rev	ATTCTTCTACTAGACAGCGAAAGAGTAGGC	
HDAC6_intr28_rev	ATATGGTCTGCCACCATGAAGCCTCTGAA	
MAPK13_inr_rev	GTGAAACAGGTCAGGGTCAGCAA	
RasGRP1_rev	ACATTGGTCCTTTGCAGCTTT	
CCND1_rev	AAATTAGGAAGGAGCCTATCGTGT	
GNG7_intr_rev	GTCACGGTGCTGTAGGTCATA	
TP$^-$ Gene block sequence	GACCATCCTTTGCGGCCGCTAACTGACATGGCCCGGAGCACCCTGAGCAAGCCTCTTAAGAACAAAGTGAATCCCCGGGGACCTCTGATCCCCTTAATTCTTCTGATGCTCAGAGGGGTCAGTACTGCTTCGCCCGG	
TP$^-$ gene block cloning primers		
NCACXN_ScaI6330_rev	CCGGGCGAAGCAGTACTGA	
NCACXN_NotI6220_fwd	GACCATCCTTTGCGGCCG	
NCAC_8290_rev	GGCGTTACTTAAGCTAGCTTGCC	
NCACXN_6327_fwd	AGTACTGCTTCGCCCGGCT	
Mouse DNA primers		
Mouse_mt_COX2_fwd	TTCTACCAGCTGTAATCCTTA	(143)

Mouse_mt_COX2_rev	GTTTTAGGTCGTTTGTTGGGAT	(143)
Mouse_IAP_fwd	ATAATCTGCGCATGAGCCAAGG	(142)
Mouse_IAP_rev	AGGAAGAACACCACAGACCAG	(142)
CCD Mutants primers		
102510NdeIINteinIN forward	ggaattccatatgATAGAAAATTCATCACCCTACACCTCAG	
102510XhoIInteinIN1-407 reverse	ccggctcgaGGGCCTCGCGGGTTAACC	
E266A_fwd	CCCATGGCCTCACCCCATATGCCATCTTATATGGGGCACCCCCGCC	
E266A_rev	GGCGGGGGTGCCCCATATAAGATGGCATATGGGGTGAGGCCATGGG	
E266K_fwd	CCCATGGCCTCACCCCATATAAGATCTTATATGGGGCACCCCCGCC	
L268A_fwd	CATGGCCTCACCCCATATGAGATCGCCTATGGGGCACCCCCG	
L268A_rev	GCGGGGGTGCCCCATAGGCGATCTCATATGGGGTGAGGCCAT	
Y269A_fwd	GCCTCACCCCATATGAGATCTTAGCCGGGGCACCCCCGC	
Y269A_rev	GCGGGGGTGCCCCGGCTAAGATCTCATATGGGGTGAGGC	
Illumina Linkers Red – Illumina linker sequence		
Linker 1 short strand	5'-PO_4-GTCCCTTAAGCGGAG-NH_2-3'	
Linker 1 long strand	GTAATACGACTCACTATAGGGCCTCCGCTTAAGGGA	
Linker 2 short strand	5'-PO_4-CGAGGCGTCTAATGC-NH_2-3'	
Linker 2 long strand	GCTATAGCAGCACATCAGTTAGGCATTAGACGCCTCGT	
Linker 3 short strand	5'-PO_4-CTATGACGGTGACGC-NH_2-3'	
Linker 3 long strand	GAGAATCCATGAGTATGCTCACGCGTCACCGTCATAGT	
Linker 5 short strand	5'-PO_4-CTGAGACGTCGATGC-NH_2-3'	
Linker 5 long strand	GATCATGCGAGATACATCTCAGGCATCGACGTCTCAGT	
Linker 6 short strand	5'-PO_4-CGATGCGGTAACTGC-NH_2-3'	
Linker 6 long strand	GTATCTCAACAAGCAGCTTGAGGCAGTTACCGCATCGT	
Linker 7 short strand	5'-PO_4-CTAGTACGGAGTCGC-NH_2-3'	
Linker 7 long strand	GCCATGGAATATGCAATCTGACGCGACTCCGTACTAGT	
Linker 8 short strand	5'-PO_4- CGGTGAGCGCATATC-NH_2-3'	
Linker 8 long strand	CAACTTGCGTGCAATTAACGAGGATATGCGCTCACCGT	

Linker 9 short strand	5'-PO$_4$- ACGTAGGTGCGCATC-NH$_2$-3'
Linker 9 long strand	CAGGATGCGTAATACGAATCTCGATGCGCACCTACGTT
Linker 10 short strand	5'-PO$_4$-CCGGTCAGCATAGTG-NH$_2$-3'
Linker 10 long strand	GACTTGAACCGTAGCATCTAAGCACTATGCTGACCGGT
Linker 11 short strand	5'-PO$_4$- CTGATACCGGCGTAG-NH$_2$-3'
Linker 11 long strand	GAGCCTACGTTACGCAATATAGCTACGCCGGTATCAGT
Linkers specific primers Blue – adapter sequence Green – sequencing primer binding site Red – Illumina linker sequence	
Linker 1	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGTAATACGACTCACTATAGGGC
Linker 2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCTATAGCAGCACATCAGTTAG
Linker 3	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAGAATCCATGAGTATGCTCAC
Linker 5	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGATCATGCGAGATACATCTCAG
Linker 6	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGTATCTCAACAAGCAGCTTGAG
Linker 7	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCCATGGAATATGCAATCTGAC
Linker 8	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCAACTTGCGTGCAATTAACGAG
Linker 9	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCAGGATGCGTAATACGAATCTC
Linker 10	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGACTTGAACCGTAGCATCTAAG
Linker 11	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAGCCTACGTTACGCAATATAG
MLV LTR U3 second round PCR Blue – adapter sequence Green – sequencing primer binding site Red – unique barcode or index	

MLV U3 LTR Index Primer - ACTTGA	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACTTGA**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index Primer - GATCAG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GATCAG**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index Primer - GGCTAC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GGCTAC**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index Primer - TAGCTT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**TAGCTT**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index primer - GTGGCC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GTGGCC**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index primer- CGTACG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGTACG**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index primer - GAGTGG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GAGTGG**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index primer - ACTGAT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACTGAT**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index primer - CTTGTA	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**CTTGTA**CCAAACCTACAGGTGGGGTCTTTC
MLV U3 LTR Index primer - ATTCCT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**ATTCCT**CCAAACCTACAGGTGGGGTCTTTC
MLV LTR U5 second round PCR Blue – adapter sequence Green – sequencing primer binding site Red – unique barcode or index	
MLV U5 LTR Index Primer - ACTTGA	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACTTGA**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index Primer - GATCAG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GATCAG**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index Primer - GGCTAC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GGCTAC**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index Primer - TAGCTT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**TAGCTT**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index primer - GTGGCC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GTGGCC**TGACTACCCGTCAGCGGGGGTC

MLV U5 LTR Index primer- CGTACG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGTACG**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index primer - GAGTGG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**GAGTGG**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index primer - ACTGAT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACTGAT**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index primer - CTTGTA	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**CTTGTA**TGACTACCCGTCAGCGGGGGTC
MLV U5 LTR Index primer - ATTCCT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT**ATTCCT**TGACTACCCGTCAGCGGGGGTC

[1] Described as primer 6320 in (135); lower case letters are not encoded by virus

[2] Described as primer 3807 in (135); lower case letters are not encoded by virus

Experimental Results

Section I. Effects of disrupting MLV IN:BET interaction on tumorigenesis in *MYC/Runx2* mouse model

The strong integration preference of MLV towards transcriptionally active regions is a consequence of IN:BET protein interactions. The primary known interaction of MLV with BET protein is between MLV IN CTD TP and ET domain of a BET protein. A previous study (154) suggested that there are residues in CCD (E266, L268, Y269) that potentially interacts with BET protein. Alanine mutations of these residues resulted in loss of interactions with Brd2 protein by co-immunoprecipitation experiments. The results presented investigates these CCD residues structurally and the effect of alanine mutations on viral pathogenesis. Finally, the effects of virus lacking the IN TP were examined using the *MYC/Runx2* mouse model.

Structural analysis of MLV IN CCD

The structure of a MLV CCD dimer has not been experimentally defined, although PFV based molecular model has been generated (59) (Fig 7A and 7B). Knowledge of the structure of the CCD dimer and, in particular the dimer interface, would provide insights on the importance of these residues (E266, L268, Y269) on the structural integrity of IN. Initial research focused on expressing the MLV CCD in bacterial expression vectors. Some problems encountered in resolving the structure of CCD was the protein solubility during purification and the protein yield. Purification of MLV IN CCD was optimized with a construct, MLV IN CCD (residues 106-328) with a C209S mutation. The C209S mutation removes the odd number of cysteine and improved protein yield compared to previous purifications performed. Addition of a zwitterionic detergent, Zwittergent 3-12, with a low concentration of 0.01% help improved solubility of MLV CCD IN in low salt conditions and can be concentrated up to 7.6 mg/mL (Fig 8A). This protein

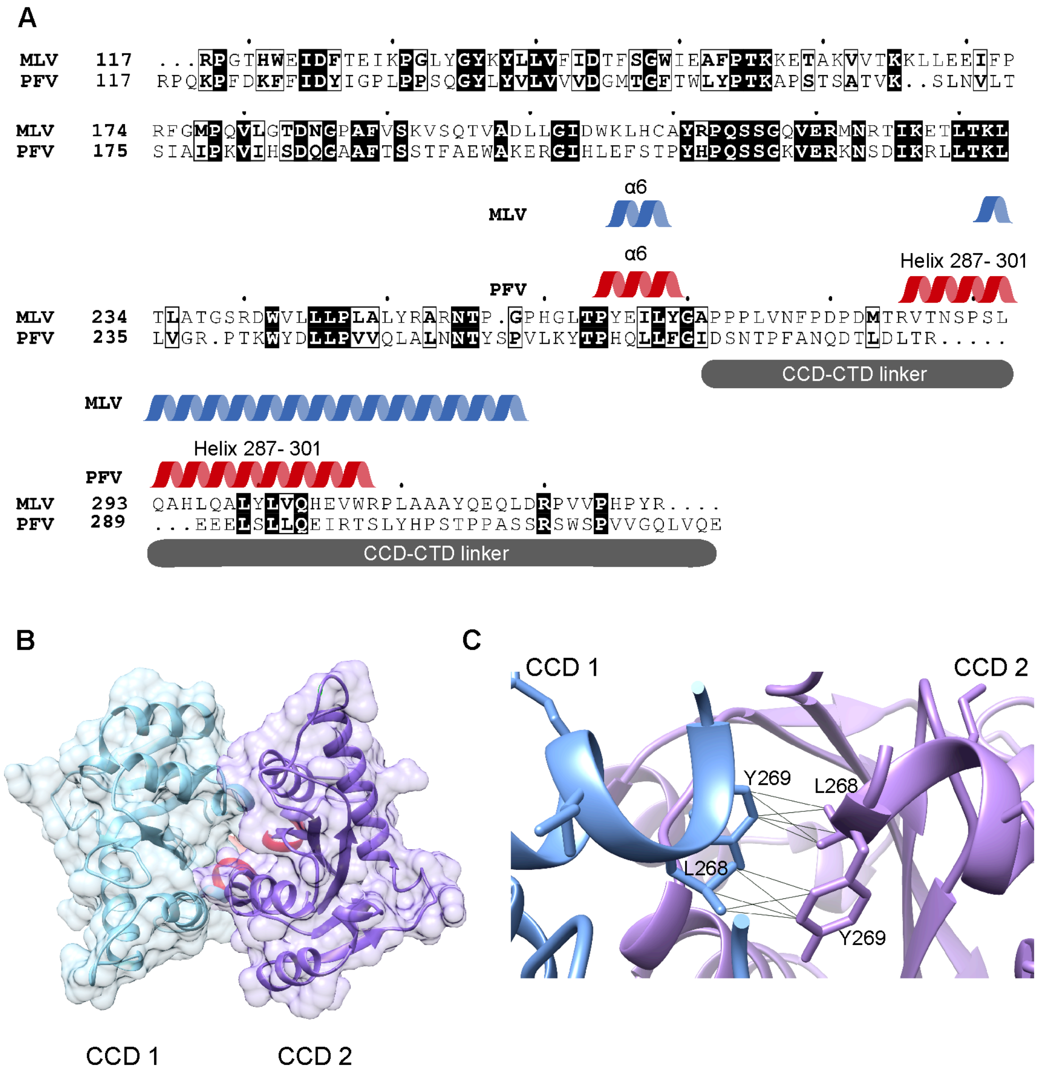

A
MLV 117 ...RPGTHWEIDFTEIKPGLYGYKYLLVFIDTFSGWIEAFPTKKETAKVVTKKLLEEIFP
PFV 117 RPQKPFDKFFIDYIGPLPPSQGYLYVLVVVDGMTGFTWLYPTKAPSTSATVK..SLNVLT
MLV 174 RFGMPQVLGTDNGPAFVSKVSQTVADLLGIDWKLHCAYRPQSSGQVERMNRTIKETLTKL
PFV 175 SIAIPKVIHSDQGAAFTSSTFAEWAKERGIHLEFSTPYHPQSSGKVERKNSDIKRLLTKL
α6
MLV
α6
Helix 287- 301
PFV
MLV 234 TLATGSRDWVLLLPLALYRARNTP.GPHGLTPYEILYGAPPPLVNFPDPDMTRVTNSPSL
PFV 235 LVGR.PTKWYDLLPVVQLALNNTYSPVLKYTPHQLLFGIDSNTPFANQDTLDLTR.....
CCD-CTD linker
MLV
Helix 287- 301
PFV
MLV 293 QAHLQALYLVQHEVWRPLAAAYQEQLDRPVVPHPYR....
PFV 289 ...EEELSLLQEIRTSLYHPSTPPASSRSWSPVVGQLVQE
CCD-CTD linker
B
CCD 1
CCD 2
C
CCD 1
CCD 2
Y269
L268
L268
Y269

Fig 7. MLV IN CCD and PFV IN CCD. (A) Structure based sequence alignment of PFV IN CCD and MLV IN CCD with secondary structure prediction using PROMALS3D (155) and displayed using ESPript (156). PFV secondary structures (red helices) are derived from PFV intasome structure (3OS1). MLV secondary structures (blue helices) are assigned using the homology model of CCD. (B) Homology model of MLV IN CCD (residues 117-271) dimer was aligned using PFV intasome (3OS1) (59). Residues 266-269 (EILY) in the α6 helix are in red. (C) Pseudobonds (black) between residues L268 and Y269 were predicted using UCSF Chimera (157).

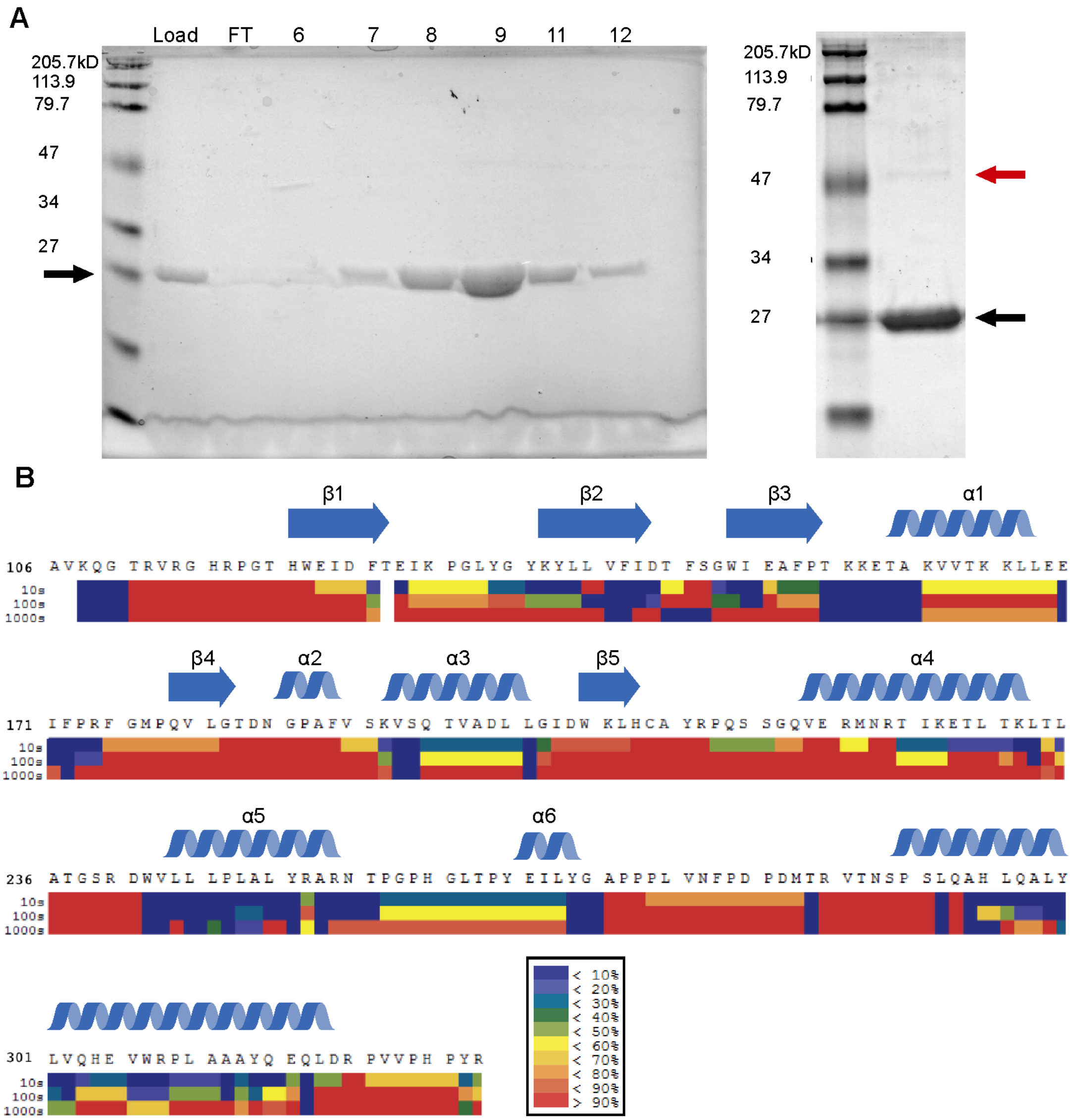


Fig 8. Hydrogen deuterium exchange mass spectrometry of MLV IN CCD$_{106-328}$ C209S. (A) MLV IN CCD was purified using cation exchange chromatography (HiTrap™ SP HP) (left panel). The fractions are labeled on top and the black arrows signify MLV IN CCD (MW: ~ 25 kDa). Purified MLV IN CCD is concentrated and 10 µg is visualized in SDS-PAG. (B) The percentage hydrogen deuterium exchange shown at 10, 100 and 1000 seconds time points. The heat map legend (in a black box) corresponds to hydrogen deuterium exchange. Secondary structures (blue) predicted in the molecular model (59) located on top for each row protein sequence.

is a dimer in solution based on analytical gel filtration (AGF) (data not shown). Based on the molecular model of the CCD dimer, interaction of Y269 and L268 were predicted at the dimer interface (Fig 7B).

A solution-based study using hydrogen deuterium exchange and mass spectrometry (HDX-MS) was performed on MLV IN CCD$_{106-328}$ C209S to define structured regions of the protein dimer. The qualitative deuterium exchange pattern showed high exchange for most of the protein including regions with predicted secondary structures (Fig 8B). It can be indicative of low quality HDX-MS analysis. The residues EILY$_{266-269}$ have 30-80% deuterium exchange over the time course (10s-1000s), which implies that this region is more structured than regions with 100% exchange after 10s (Fig 8B). Furthermore, these residues are located in the predicted α-6 of MLV CCD dimer. This is supportive of the possible interaction between Y269 and L268 at the dimer interface (Fig 7B).

Mutational analysis of MLV IN CCD residues implicated with BET interaction

It was of interest to further examine the effects of these putative IN CCD mutations proposed to be involved in IN:BET protein interactions on integration. However, MLV bearing IN E266A, IN L268A, and IN Y269A were equivalent to the catalytically inactive IN D184N active site mutant (Fig 9A). Thus, tissue culture and animal studies with these viruses are not possible. These mutations were also introduced into full length IN in a pET15 NESG to observe effects in protein multimerization but experiments were not performed.

Initial study on M-MLV replication competent IN XN virus

Previous *in vitro* studies of MLV virus bearing truncation of the IN CTD TP indicated these viruses had decreased bias towards integration into CpG islands and TSS (158), due to the loss of interaction with the host BET proteins.

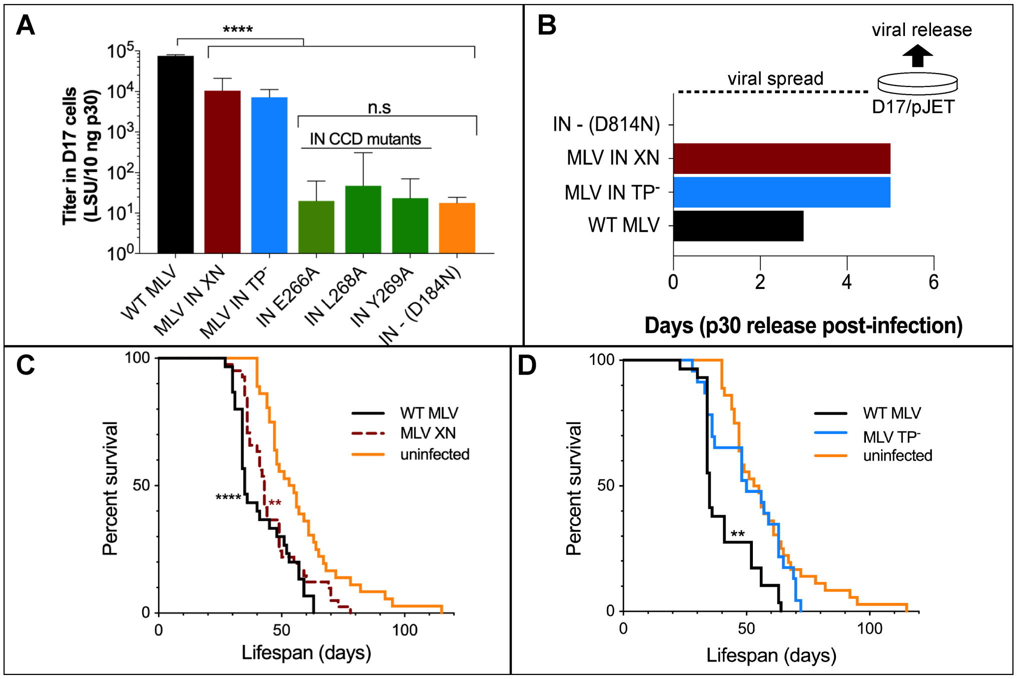


Fig 9. Viral characteristics of MLV IN TP$^-$ in cell culture and in the *MYC/Runx2* mice model. (A) LacZ titers of the various IN mutants constructs: WT MLV (black), MLV IN-XN (dark red), MLV IN-TP$^-$ (blue), IN CCD mutants (green), and IN - D184N (orange). Dunnett's Multiple comparison test: ****P<0.0001, n.s = no significance. Error bars indicate SEM; n=3. (B) Viral spread of MLV IN mutants and WT MLV in D17/pJET measured by p30 (CA) released into media. Proviral DNA was transiently introduced into cells using DEAE dextran. Viral supernatants were collected at the indicated days and levels of CA were detected by ELISA (132). (C) Survival curves of *MYC/Runx2* mice infected neonatally with MLV IN-XN. WT MLV (solid black, MLV IN-XN (dashed brown), uninfected (solid orange). Log-rank test survival curve comparisons: MLV WT (n=30) vs. MLV IN-XN (n=40) P=0.0889; MLV WT vs. uninfected (n=36), ***P<0.0001; MLV IN-XN vs. uninfected, **P≤0.0021 (G) Survival curves of *MYC/Runx2* with infected neonatally with MLV IN TP$^-$. WT MLV (black, MLV IN TP$^-$ (blue), uninfected (orange). Log-rank test survival curve comparisons: MLV WT (n=30) vs. uninfected (n=36) = ****P<0.0001; MLV WT vs. MLV IN TP$^-$ (n=23), *P=0.0061; MLV IN TP$^-$ vs. uninfected, P=0.2632

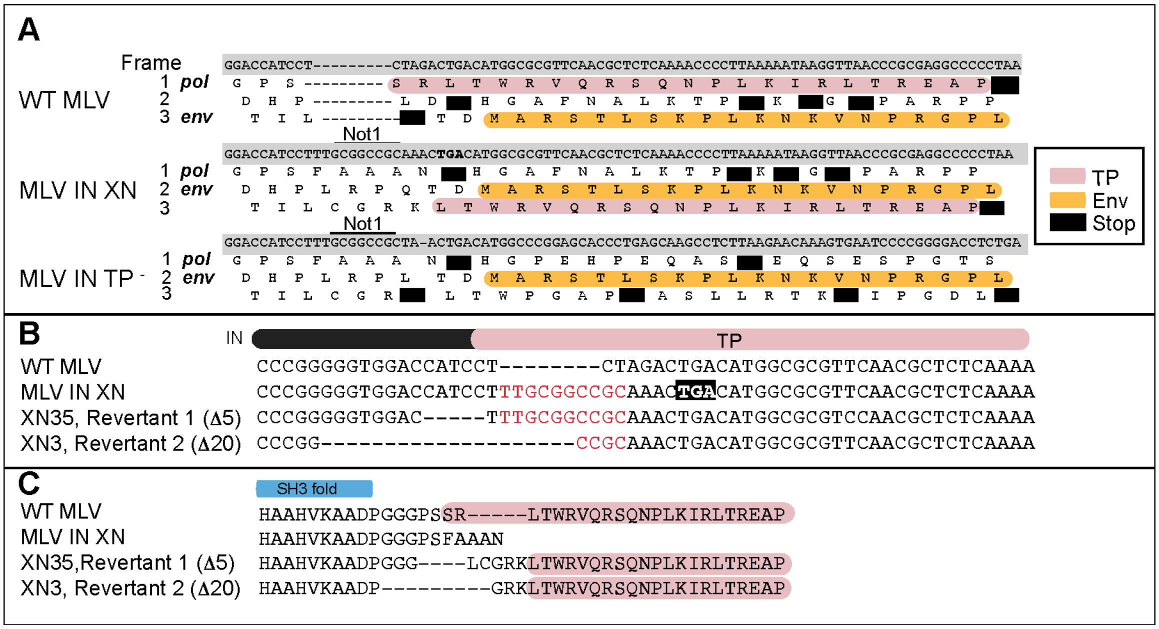

A
Frame
WT MLV
1 pol
2
3 env
GGACCATCCT---------CTAGACTGACATGGCGCGTTCAACGCTCTCAAAACCCCTTAAAAATAAGGTTAACCCGCGAGGCCCCCTAA
G P S --------- S R L T W R V Q R S Q N P L K I R L T R E A P
D H P --------- L D H G A F N A L K T P K G P A R P P
T I L --------- T D M A R S T L S K P L K N K V N P R G P L
Not1
MLV IN XN
1 pol
2 env
3
GGACCATCCTTTGCGGCCGCAAACTGACATGGCGCGTTCAACGCTCTCAAAACCCCTTAAAAATAAGGTTAACCCGCGAGGCCCCCTAA
G P S F A A A N H G A F N A L K T P K G P A R P P
D H P L R P Q T D M A R S T L S K P L K N K V N P R G P L
T I L C G R K L T W R V Q R S Q N P L K I R L T R E A P
Not1
MLV IN TP -
1 pol
2 env
3
GGACCATCCTTTGCGGCCGCTA-ACTGACATGGCCCGGAGCACCCTGAGCAAGCCTCTTAAGAACAAAGTGAATCCCCGGGGACCTCTGA
G P S F A A A N H G P E H P E Q A S E Q S E S P G T S
D H P L R P L T D M A R S T L S K P L K N K V N P R G P L
T I L C G R L T W P G A P A S L L R T K I P G D L
TP
Env
Stop
B
IN
TP
WT MLV CCCGGGGGTGGACCATCCT--------CTAGACTGACATGGCGCGTTCAACGCTCTCAAAA
MLV IN XN CCCGGGGGTGGACCATCCTTTGCGGCCGCAAACTGACATGGCGCGTTCAACGCTCTCAAAA
XN35, Revertant 1 (Δ5) CCCGGGGGTGGAC-----TTTGCGGCCGCAAACTGACATGGCGCGTCCAACGCTCTCAAAA
XN3, Revertant 2 (Δ20) CCCGG--------------------CCGCAAACTGACATGGCGCGTTCAACGCTCTCAAAA
C
SH3 fold
WT MLV HAAHVKAADPGGGPSSR-----LTWRVQRSQNPLKIRLTREAP
MLV IN XN HAAHVKAADPGGGPSFAAAN
XN35,Revertant 1 (Δ5) HAAHVKAADPGGG----LCGRKLTWRVQRSQNPLKIRLTREAP
XN3, Revertant 2 (Δ20) HAAHVKAADP---------GRKLTWRVQRSQNPLKIRLTREAP

Fig 10. MLV IN without TP$^-$ constructs and viral revertants. (A) Alignment of the three MLV IN constructs, WT MLV, MLV IN-XN and MLV IN TP$^-$ in the overlap region of IN TP (pink) and Env (orange) reading frames. Black boxes represent stop codons. (B) Viral revertants identified in *MYC/Runx2* tumors infected with MLV IN-XN. 3' terminus of IN is aligned with the region corresponding to the IN (black bar) and IN TP (pink) indicated. Not1 linker insertion (129) generating the IN-XN is indicated in red. Deletions (Δ5 and Δ20) with respect to IN-XN are indicated by dash lines. Premature TGA stop codon for IN-XN is shown with the black box. (C) Protein alignment of WT MLV, MLV IN-XN and IN revertants. Deletions were localized between the SH3 fold (blue) and the TP (pink) of IN.

The MLV IN TP region is overlapping with the *env* coding sequences in alternative open reading frames. In order to analyze the effects of M-MLV virus bearing truncations of the IN C-terminus in the mouse model, replication competent viruses were needed that terminated the IN protein without altering the expression of the ecotropic M-MLV Env. Initial experiments were performed using the pNCA-C IN-XN construct (129), which truncated the C-terminal 23 aa and maintained a viral titer in 293T cells within 2 fold of WT IN (60). The IN-XN construct was generated by a linker-insertion that creates a stop codon within IN upstream of the *env* coding region and results in a frame-shift of the sequence encoding the IN C-terminal region (Fig 10). The effects of virus lacking the BET interaction domain within the IN TP were examined within the mouse *MYC/Runx2* model. This model is beneficial, in that lymphomas form within 36 days, and are accelerated by the insertional mutagenesis of M-MLV at known CIS by approximately 10 days (90). Survival experiments in the *MYC/Runx2* mouse model, infecting with WT M-MLV, MLV IN-XN and comparing to an uninfected control, were performed by injecting MLV virus neonatally and then recording the Date of Death (DoD) over a 115-day period (Appendix 1). The survival curves of WT (n=30) and IN-XN (n=40) mice did not show statistical significance to each other (P=0.089) using the Log-rank test test but both showed significantly significant differences using the same test to the uninfected control (****P<0.0001, **P<0.0021 respectively) (Fig 9C). Median survival times of WT and IN-XN virus were 35 and 43 days, respectively, and both succumbed to tumors significantly faster than the uninfected control (median survival 54 days). To investigate further, IN-XN virus isolated from the tumors of three infected mice was introduced into 293mCAT cells, human cells expressing the mouse ecotropic receptor (30), which facilitated the isolation of infectious virus in the absence of endogenous mouse viruses. Remarkably, the viruses transferred to 293mCAT cells from two independent mice (XN3; DoD 30d and XN35; DoD 70d), harbored deletions of 20 and 5 bases, respectively (Fig 10B). These deletions removed the IN stop codon and restored the IN C-terminus, encoded in an alternate reading frame, onto the IN protein (Fig 10C). Functionally, such deletions alter the spacing between the IN CTD SH3 fold and the region of the C-terminus that becomes structured upon binding to the host Brd ET domain (Fig 10C) (60). DNA from the XN2 mouse (DoD=41d) maintained the IN-XN genotype.

Generation and characterization of M-MLV replication competent IN TP$^-$ virus

In order to eliminate the possibility of small deletions restoring the IN tail peptide onto the IN protein, a second construct, pNCA-C TP$^-$. The construct design eliminated the coding potential of IN TP within the *env* overlapping region and incorporated multiple stop codons into the non-*env* reading frames (Fig 10A). Single round infection into D17 cells for *lacZ* transfer confirmed that the MLV IN TP$^-$ produces the same viral titer as MLV IN-XN, showing in this assay, a 10-fold decrease compared with the wild-type (WT) MLV (Fig 9A). Titers of both MLV IN-XN and TP$^-$ were 10^3 fold higher than the IN mutant in the catalytic triad (D184N) (130).

Viral passage of IN-XN and TP$^-$ in D17/pJET, expressing the ecotropic mCAT receptor (127) displayed similar time course of infection, with viral CA detected in the media at day 5, a delay of 2 days from the WT MLV (Fig 9B). Viral competency was further confirmed by western blot analysis of CA, IN, Env proteins associated with viral particles released from D17/pJET viral producer cell lines (Fig 11). Env (74 kDa) and CA (30 kDa) are expressed in both MLV IN TP$^-$ and WT MLV viral particles. The truncation of the IN TP was stable after passage in tissue culture (Fig 11, lane 1), maintaining the predicted decreased molecular weight compared with WT IN (Fig 11, lane 2). No viral proteins (Env, CA and IN) were detected in the replication defective pNCA-C IN D184N virus (Fig 11, lane 3).

Having verified that the MLV IN TP$^-$ virus was replication competent and that the truncation was stable, survival experiments were performed in *Myc/Runx2* mice to determine if this optimized construct affected tumorigenesis (Fig 9D) (Appendix 2). As previously reported (141), the mice infected with WT MLV exhibited significantly poorer survival than the uninfected controls (Log-rank test: ****P<0.0001; Fig 1G). Interestingly, the survival curve of MLV IN TP$^-$ infected mice (n=23) showed no significant differences (P=0.2632) from that of the uninfected mice (n=36), in contrast to the observed curve for the MLV IN XN virus (Fig 9C). Of note, the survival curve for IN TP$^-$ shows a biphasic trend, with one third of the mice displaying early tumors. Additionally, all of the IN TP$^-$ mice were deceased by day 72, while the uninfected mice survived until day 115. The median survival for mice infected with WT MLV or MLV IN TP$^-$ and uninfected mice was 35, 50, and 53 days, respectively. These observed phenotypical differences suggest that MLV IN TP$^-$ delayed tumorigenesis in the mice model.

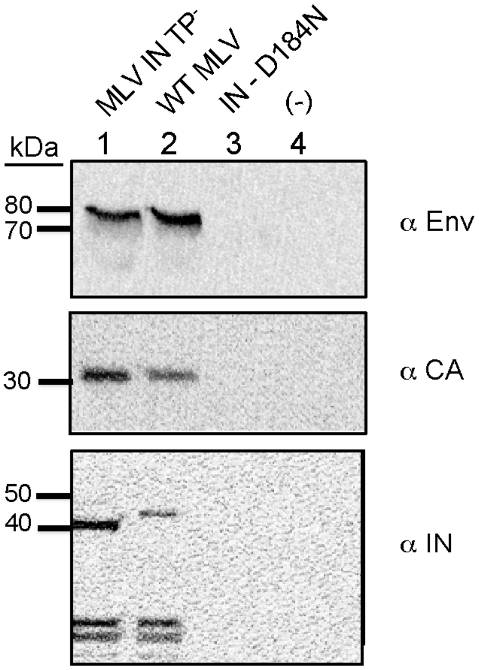


Fig 11. Western blot analysis of MLV virus-associated proteins. DNA of proviral DNA constructs (pNCA-C) encoding WT, IN TP$^-$ and IN D184N were transiently introduced into D17/pJET cells using DEAE dextran, passaged for 14 days to allow viral spread and viral supernatants were collected, pelleted by centrifugation and analyzed by PAG and western blot using anti-SU (80S-019), anti-CA (81S-263) and anti-IN antibodies (136). Supernatants from D17/pJET cells were used as a negative control. Positions of the protein standards are indicated at the left. Predicted MW of the WT MLV viral proteins are SU (75 kDa), CA (30 kDa) and IN (45 kDa).

Discussion

The discovery that the MLV IN protein interacts with host BET proteins, directing integration into sites of active expression marked by acetylated chromatin, gave rise to the question of whether the loss of the MLV BET binding site would affect the pathogenesis of the virus. Co-immunoprecipitation studies with IN CCD mutations included it as a potential BET secondary interaction where viral competency hasn't been analyzed.

The possible role IN CCD in BET interaction was investigated in this study specifically the three amino acids E266, L268, and Y269 described previously (154). Molecular modeling indicates that these positions are located in the predicted MLV CCD α6 helix (264-270) and are close to the CCD dimer interface (Fig 7B). The model shows close proximity of the Y269 and L268 between monomers and thus alanine mutations of L268, and Y269 could result in disruption of CCD multimerization. HDX analysis showed that these residues are in fact in a relatively structured region of CCD dimer therefore could be evidence that these residues are involved in dimer interface of CCD.

A common problem encountered in HDX-MS is the loss of the deuterium labeling or back-exchange in H_2O during sample preparation. If the quenching of H/D exchange is not efficient, sites that have taken up deuterium will exchange with H_2O during the proteolysis and peptide separation phases of sample preparation. In this case, dynamic regions will be artificially mislabeled with low deuterium exchange and interpreted as structured region (140, 159). This is not what we observe for the qualitative deuterium exchange pattern of IN CCD C209S HDX-MS analysis. Such back-exchange can't explain high deuterium exchange in predicted structured regions (blue arrows, Fig 8B). Possible other explanations for the unexpectedly high amount of amide proton exchange in this construct include improper quenching or sample handling, resulting in extensive exchange to deuterium before proteolysis (140, 160).

Alternatively, the protein could in fact be partly unfolded once the detergent concentration was reduced for the HDX-MS experiment. Both scenarios could cause apparent high or complete exchange at 10 seconds in polypeptide regions with predicted secondary structures. This high exchange of deuterium observed in polypeptide regions which we expect to be well-ordered reduces our confidence in the reliability of this HDX-MS analysis. Repeating this experiment would provide a more definitive evidence of the structural dynamics of the MLV CCD dimer.

We have analyzed these mutants in the context of single-round infection using *lacZ* transfer, which indicated that all three of these individual mutants resulted in viral titers equivalent to the IN catalytic mutant D184N (Fig 9A and data not shown). Regardless of the mechanism, the lack of viral titer eliminates these constructs for analysis of target site selection in tissue culture or within the mouse model.

Disruption of the BET interaction was investigated by infection in *MYC/Runx2* mice, a transgenic model displaying rapid tumor formation that is further accelerated by MLV infection insertion (99, 100, 141), with MLV lacking the IN tail peptide. Globally, the infection time course for the IN TP$^-$ series showed a biphasic DoD curve, with 1/3 of the mice developing tumors early, as with WT infection, and 2/3 of the mice developing tumors later, paralleling the uninfected *MYC/Runx2* control mice.

These results highlight the strong bias within the mouse to maintain the presence of the IN-TP. Interestingly, the TP region in encoded within the region overlapping with the Env signal peptide, yet the codon bias within 4070A MLV is maintained towards the IN reading frame (161). Sequence conservation between gammaretroviruses identified the conserved sequence W(X$_7$)PLK(I/L)R within the TP (60). From our studies, the initial viruses (MLV IN-XN) emerging from the tumors indicated that the selective pressures restored the TP through deletion mutations within the virus coding region, removing a small putative linker region within the IN C-terminus to restore the tail peptide encoded in an alternative reading frame. When this option was removed, through optimizing codon usage for the IN/Env region towards the *env* sequence, thereby destroying the coding potential of the TP, the circulating virus detected restored the TP through recombination with the endogenous polytropic viruses, with cross-over junctions within the IN and Env coding regions.

The deletion mutations that restore the BET interacting domain onto the IN-XN construct identify a linker region between the IN CTD SH3 fold and the IN BET ET domain. The IN SH3 fold has been defined structurally by NMR to extend through the KAADPG sequence (59), which is in agreement with IN truncation studies of the MLV IN (129, 162), where truncations after KAADP are viable. The region after the SH3 fold that that includes the linker region and IN TP is considered as an intrinsic disordered region based on NMR analysis. When IN TP interacts with ET domain, it becomes structured and the linker region provides flexibility for the complex. IN W390 is required for tight binding to BET proteins (60, 70). After passage in mice, the spacing between the IN SH3 and TP domains is reduced from nine amino acids (GGGPSSRLT) to 3 amino acids (GRK), two of which are not normally encoded by the IN protein. This indicates that the length of this linker region as well as the composition of the amino acids between the IN SH3 fold and the BET binding site can be substantially altered and maintain virus viability in vivo.

Section II. Analysis of integrations and recombination in mice tumor

Analysis of WT AND IN TP$^-$ MLV integration preference at TSS and CpG islands in *MYC/Runx2 mice*

Previous studies have shown that truncation of IN TP decreases preferential integration at TSS and CpG islands by >50% in tissue culture (60, 70), and experiments were performed to determine if this was also the case in the *MYC/Runx2* mouse. Fig 12 outlines the overall workflow for analysis of IN TP$^-$ virus integration sites. WT IN and IN TP$^-$ virus integration sites were mapped from genomic DNA samples from *MYC/Runx2* mouse tumors using ligation-mediated PCR (LM-PCR) and next generation sequencing (NGS) (Table 1). Libraries were generated from tumors from 4 representative mice infected with WT MLV (WT6, 8, 10, and 12) and 5 representative mice infected with MLV IN TP$^-$ (TP$^-$4, 6, 7, 9, and 16) and integrantion sites were mapped as described in Materials and Methods (Table 1). Integration sites mapping to +/- 1 kb from TSS and CpG islands were noted, as described previously (163). Integration around TSSs and CpG islands was 11.8-13.4% and 11.9-14.3%, respectively, for WT tumors. Surprisingly, tumors from 4 of the 5 *MYC/Runx2* mice infected with MLV IN TP$^-$ that were analyzed (TP$^-$4, 6, 7, and 9) showed similar preferential integration at TSSs and CpG islands to the WT IN (10.2-11.9% and 11.3-13.4% respectively). In contrast, TP$^-$16 tumor integration at TSSs and CpG islands were markedly lower from that observed in both the other IN TP$^-$ and the WT MLV infected tumors.

The TP$^-$16 and WT mice had the same DoD (day 34), and so samples from these tumors were compared. Considering integrations at TSSs /CpG islands, TP$^-$16 integrants were statistically different from WT6 integrants (Appendix 4, Fisher's test, P< 0.001), with MLV IN TP$^-$16 displaying 7.3% and 6.3% of its integrations at TSSs and CpG islands, respectively, compared to 13.4 % and 13.5%, respectively, for WT6.

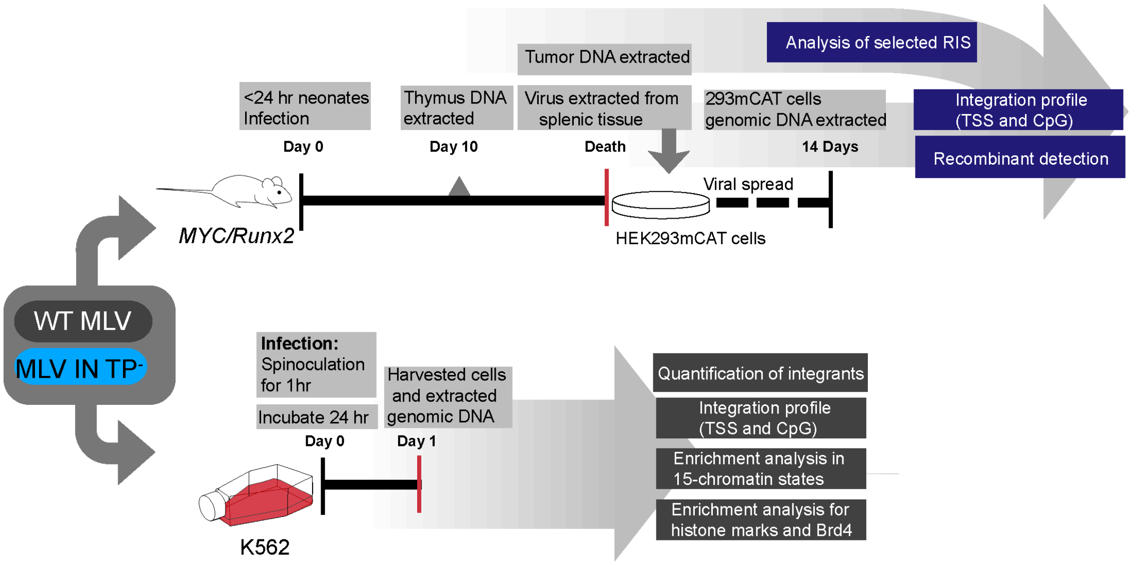

Analysis of selected RIS
Tumor DNA extracted
<24 hr neonates Infection
Thymus DNA extracted
Virus extracted from splenic tissue
293mCAT cells genomic DNA extracted
Integration profile (TSS and CpG)
Recombinant detection
Day 0
Day 10
Death
14 Days
Viral spread
MYC/Runx2
HEK293mCAT cells
WT MLV
MLV IN TP-
Infection: Spinoculation for 1hr
Incubate 24 hr
Harvested cells and extracted genomic DNA
Day 0
Day 1
K562
Quantification of integrants
Integration profile (TSS and CpG)
Enrichment analysis in 15-chromatin states
Enrichment analysis for histone marks and Brd4

Fig 12. *MYC/Runx2* mice and K562 cells studies workflow. Schematic of experiments performed with the *MYC/Runx2* mice (top panel) and K562 cells (bottom panel). Details of each experiment are found in Materials and Methods.

Table 2. MLV integration site mapping of WT MLV and MLV IN TP$^-$ tumors from *MYC/Runx2* mice

	Day of Death (DoD)	Unique sites	TSS +/- 1kb (%)	CpG +/- 1kb (%)
WT6	34	17527	2352 (13.4)	2369 (13.5)
WT8	36	4605	549 (11.9)	547 (11.9)
WT10	30	5721	678 (11.8)	697 (12.2)
WT12	50	1474	181 (12.3)	211 (14.3)
TP$^-$4	57	3565	427 (11.9)	477 (13.4)
TP$^-$6	37	2402	274 (11.4)	282 (11.7)
TP$^-$7	37	8139	912 (11.2)	986 (12.1)
TP$^-$9	63	1641	169 (10.2)	186 (11.3)
TP$^-$16	34	536	39 (7.3)	34 (6.3)

Furthermore, when comparing TP^{-}16 integrations with the mean integration level across all WT tumors (12.4% ± 0.354 and 12.9% ± 0.570 for TSSs and CpG islands, respectively), TP^{-}16 integrations at these sites decreased 1.7- and 2.0-fold , respectively, compared to WT virus. Thus, while MLV TP^{-}16 retained the expected phenotype on loss of IN TP, insertions from tumors TP^{-}4, 6, 7, and 9 were more similar to the WT virus. Unlike the IN-XN construct, the IN TP construct cannot revert to functional IN through deletions in the viral genome; however, recombination with endogenous retroviruses could have restored the IN TP in tumors from TP^{-}4, 6, 7, and 9.

Detection of recombination of M-MLV with ERVs

To verify the stability of the IN TP^{-} and to identify the cause of the TP 4, 6, 7, and 9 behaving similar to WT virus, the samples were analyzed for the presence of recombination with endogenous retroviruses (ERVs). Recombinants were detected using PCR analysis with primer pairs that include a primer for each of three different classes (amphotropic, polytropic, and xenotropic) of known ERVs and an M-MLV primer (Fig 13, Table 1, (103)). Amphotropic Env primers were used in the analysis as a control for amplification of other laboratory constructs (142).

To facilitate this analysis, infectious virus was isolated from tumor cells and used to infect human 293mCAT cells, which express the mouse ecotropic virus receptor. Polytropic, amphotropic and xenotropic MLV can also infect this cell line. Transferring the virus to 293T cells eliminates the potential for background amplification products from murine ERVs. PCR analysis of the IN/env region of MLV TP^{-}4, 6, 7, 9, 16 is shown in Fig 14, using the primers specified in Table 1. Viral IN was successfully detected by PCR in MLV TP^{-}6, 7, and 9 using an RT_universal_fwd primer and two independent polytropic-specific reverse primers (Polytropic_JS4_rev and Polytropic_JS5_rev)

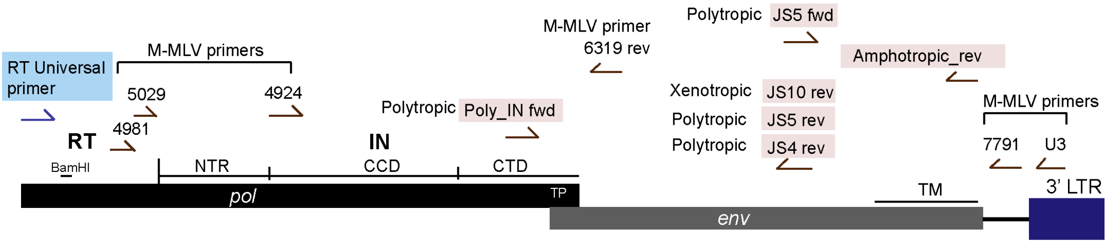


Fig 13. Scheme of the MLV genome and primers. Diagram of the 3' terminal half of the MLV genome, encoding *pol* (black line), *env* (grey line) and the LTR (blue box). Individual subdomains of the IN and Env proteins are indicated: NTR (61), N-terminal region; CCD, catalytic core domain; CTD, C-terminal domain; TP, Tail peptide; TM, transmembrane protein. Primer in blue box is a MLV universal primer located in RT, which hybridizes within a sequence conserved between known ecotropic, amphotropic, polytropic and xenotropic MLV. Primers in pink boxes are endogenous retrovirus (ERV) specific primers. M-MLV specific primers are labeled.

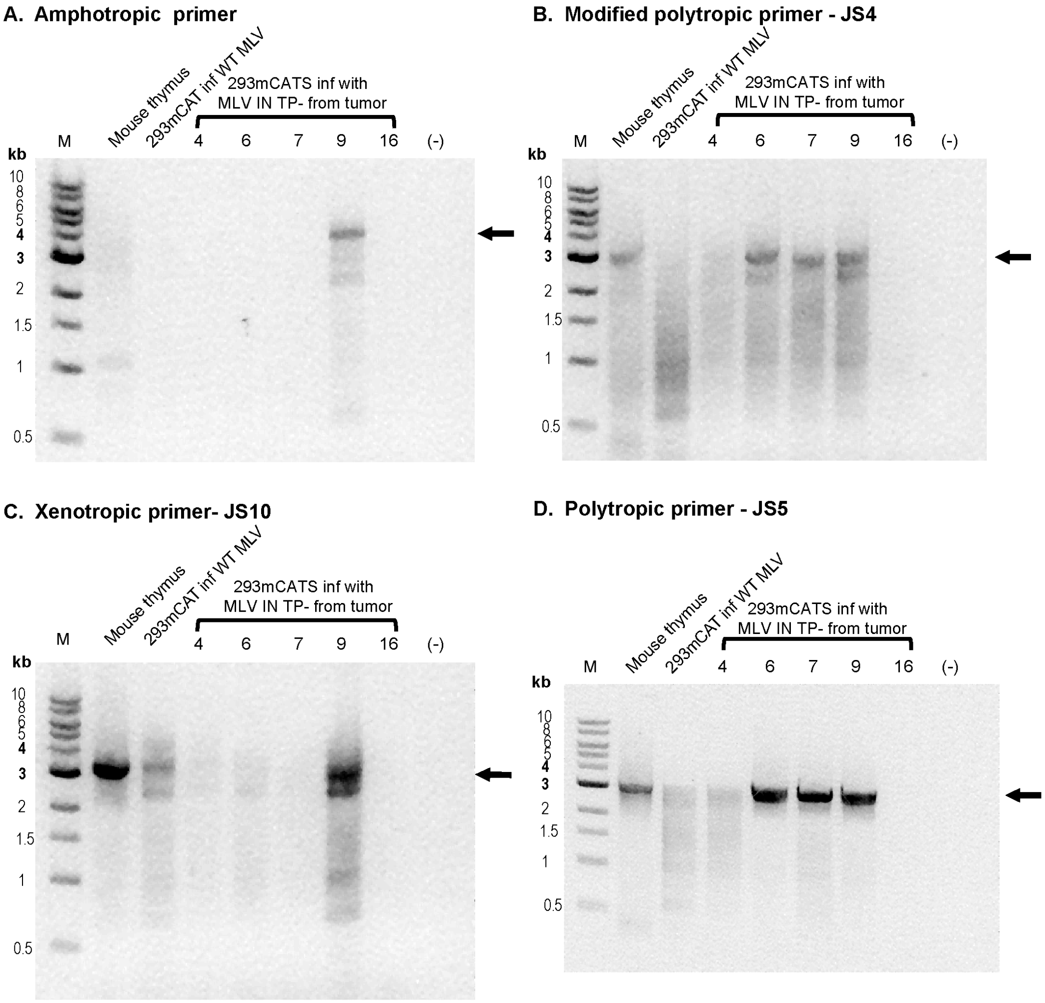


Fig 14. PCR for MLV recombinants from mice tumors. Representative agarose gels for MLV recombinant detection with universal RT primer and (A) polytropic primer (JS4) or (B) polytropic (JS5) primer. (C) Xenotropic primer (D) Ampotropic primer. These reverse primers are located in the SU or TM regions of the Env.

(Fig 13). The 293mCAT DNA were negative for mouse contamination using PCR for intracisternal particle A (IAP) and mouse mitochondrial cyclooxygenase-2cyclooxygenase-2 (mCOX2) DNA sequences (Fig 15AB). PCR using either the Amphotropic_rev or the Xenotropic_JS10_rev reverse primers in conjunction with the RT_universal_fwd primer detected only MLV TP^{-}9 (data not shown), and the quality and quantity of this product was insufficient for subsequent sequencing and analysis.

For these studies, it is of high interest to define the crossover junctions within recombinants, as they directly address the restoration of the IN TP and the receptor recognition of the subsequent virus. To further analyze the M-MLV/polytropic ERV recombinants extracted from mice TP^{-}6, 7, and 9, TA cloning was used to determine the unique 5' junction point for each virus (Fig 16A-B). The 3' junction could be determined by sequencing the PCR products described above using Polytropic_JS5_forward primer and 7791_reverse primer. The polytropic ERV (P-ERV) DNA segments from TP-6, 7 and 9 had close homology to the P-ERV *Pmv20 (110)*. The 5' junction points for the recombinants were found to be within the IN region, and the 3' junction points were in the C-terminus of the surface (SU) portion of the *Env* gene. In sum, the segments that recombined into the M-MLV virus contain the WT IN TP sequence, resulting in the expression of full-length IN during viral spread in the mice.

Having shown that recombination had occurred in some of the mice infected with MLV IN TP^{-}, tumors from other mice with early DoD were investigated to see if recombination events had occurred. Tumor DNA samples from MLV IN TP^{-}4, 12, 13, 15, 16, 17, 18, and 19 were thus analyzed by nested PCR. The primary PCR product was generated using a forward primer that hybridized to all MLV classes (RT_universal_fwd) and a reverse M-MLV primer (MLV_LTR_U3_rev) that amplifies 3' M-MLV DNA sequences including IN and Env. For the second round, polytropic recombinants were detected using a Polytropic_JS5_fwd primer and an M-MLV reverse primer (7791_rev)

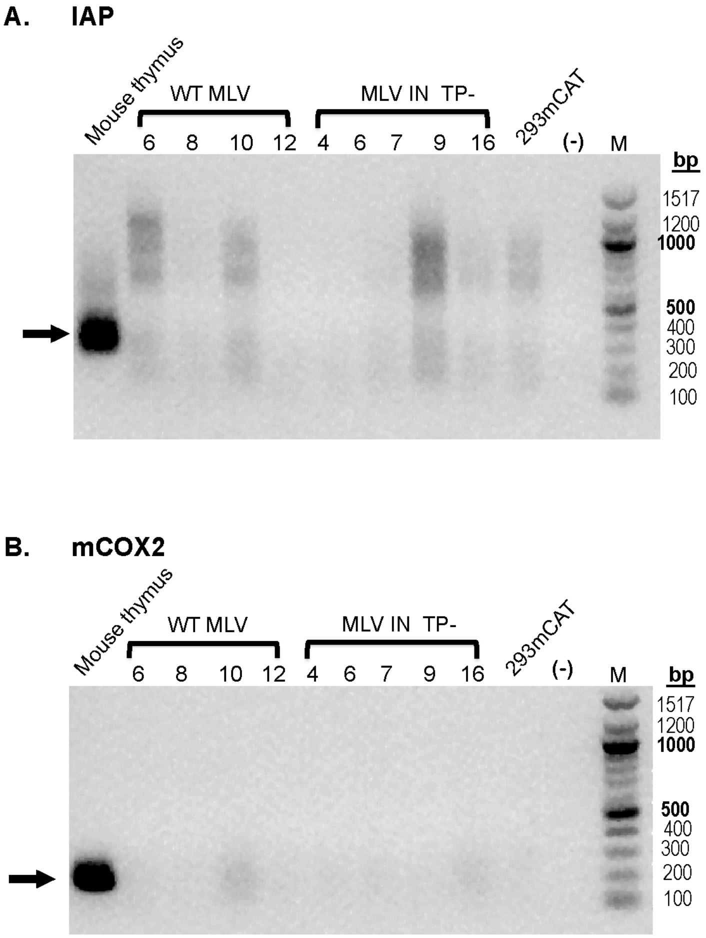


Fig 15. Detection of mouse DNA contamination in viral infected human cell line 293mCAT. (A) PCR for mouse IAP. (B) PCR for mouse COX2. Mouse thymus DNA was used as positive control. Black arrows indicate correct product size.

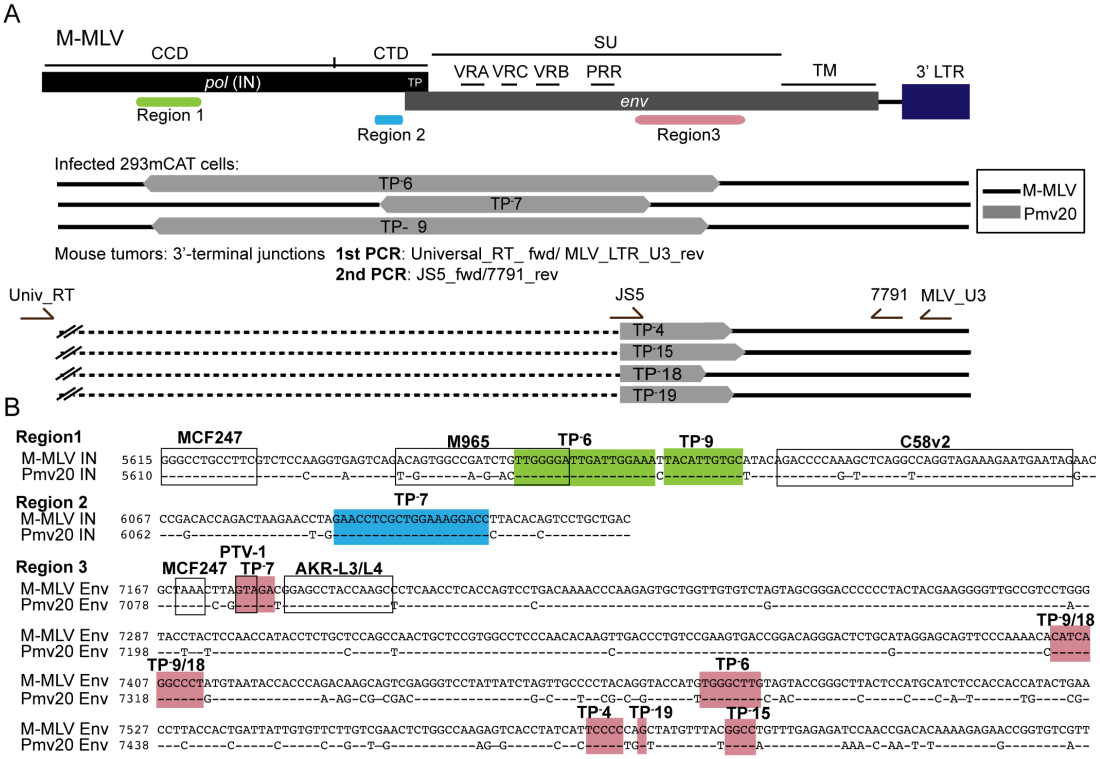

A
M-MLV
CCD
CTD
SU
pol (IN)
TP
VRA VRC VRB
PRR
TM
3' LTR
env
Region 1
Region 2
Region3
Infected 293mCAT cells:
TP-6
TP-7
TP- 9
M-MLV
Pmv20
Mouse tumors: 3'-terminal junctions
1st PCR: Universal_RT_ fwd/ MLV_LTR_U3_rev
2nd PCR: JS5_fwd/7791_rev
Univ_RT
JS5
7791
MLV_U3
TP-4
TP-15
TP-18
TP-19
B
Region1
MCF247
M965
TP-6
TP-9
C58v2
M-MLV IN 5615 GGGCCTGCCTTCGTCTCCAAGGTGAGTCAGACAGTGGCCGATCTGTTGGGGATTGATTGGAAATTACATTGTGCATACAGACCCCAAAGCTCAGGCCAGGTAGAAAGAATGAATAGAAC
Pmv20 IN 5610
Region 2
TP-7
M-MLV IN 6067 CCGACACCAGACTAAGAACCTAGAACCTCGCTGGAAAGGACCTTACACAGTCCTGCTGAC
Pmv20 IN 6062
Region 3
PTV-1
MCF247
TP-7
AKR-L3/L4
M-MLV Env 7167 GCTAAACTTAGTAGACGGAGCCTACCAAGCCCTCAACCTCACCAGTCCTGACAAAACCCAAGAGTGCTGGTTGTGTCTAGTAGCGGGACCCCCCTACTACGAAGGGGTTGCCGTCCTGGG
Pmv20 Env 7078
TP-9/18
M-MLV Env 7287 TACCTACTCCAACCATACCTCTGCTCCAGCCAACTGCTCCGTGGCCTCCCAACACAAGTTGACCCTGTCCGAAGTGACCGGACAGGGACTCTGCATAGGAGCAGTTCCCAAAACACATCA
Pmv20 Env 7198
TP-9/18
TP-6
M-MLV Env 7407 GGCCCTATGTAATACCACCCAGACAAGCAGTCGAGGGTCCTATTATCTAGTTGCCCCTACAGGTACCATGTGGGCTTGTAGTACCGGGCTTACTCCATGCATCTCCACCACCATACTGAA
Pmv20 Env 7318
TP-4
TP-19
TP-15
M-MLV Env 7527 CCTTACCACTGATTATTGTGTTCTTGTCGAACTCTGGCCAAGAGTCACCTATCATTCCCCCAGCTATGTTTACGGCCTGTTTGAGAGATCCAACCGACACAAAAGAGAACCGGTGTCGTT
Pmv20 Env 7438

Fig 16. Recombinants in MLV IN TP$^-$16 tumors and 293mCAT cells infected with tumor derived viruses. (A) Diagram of the breakpoints of the recombinants in *pol* and *env* region. The three regions where breakpoints were localized are indicated; region 1 (green), region 2 (blue) and region 3 (salmon). Recombinants isolated from mouse tumors and infected 293mCAT cells are grouped as indicated. Segments with homology to *Pmv20* are indicated in grey and M-MLV is indicated in black. For infected 293mCAT cells, primers used to identify the recombinants are in Fig13. For recombinants identified in tumors, the nested PCR primers used to identify the 3' breakpoints are shown in the diagram. The dashed black line indicates undetermined 5' junction point for those revertants. (B) The breakpoints of recombinants on the alignment of the M-MLV and *Pmv20* are shown and arranged by region. Previously reported recombinants (110) are indicated in black boxes in region 1 and region 3 (PTV-1). Coloring of the three regions are as indicated in panel 16A. Crossover regions within individual tumors are labeled.

(Fig 16A, bottom). Recombinant viruses were detected in tumor samples TP$^-$4, 15, 18 and 19, and were successfully sequenced and analyzed for 3' junction points (Fig 16B). However, 5' junctions were not identified in these recombinants. As above, the polytropic segment of these recombinants was homologous to *Pmv20*. The identified 3' junction points for the TP$^-$ recombinants were variously distributed throughout the C-terminus of the *Env* SU region, as shown in Fig 16B (region 3; bottom). The 3' junctions of TP$^-$9 and TP$^-$18 were indistinguishable in this analysis.

The recombinants identified within the IN (Region 1) had junction points unique from the MCF247, M965 and C58v2 isolates identified previously (Fig 16B, black boxes) (110). Similarly, within Region 3, the crossover for TP-7 is related to the cross-over junction identified for PTV-1 (110). However, the presence of recombination did not correlate with tumorigenesis in the *MYC/Runx2* mouse. This is exemplified when comparing the TP$^-$16 tumor, which had a DoD of 34 days and no detectable recombinants in any of its DNA samples, with TP$^-$4, 6 and 9, which all had recombinant virus and DoDs of 50, 57 and 63 days, respectively.

The integration profile of MLV IN TP$^-$16 is distinct from that of MLV WT.

Analysis of integrated MLV IN TP$^-$16 within tumor DNA did not detect recombination with ERVs and indicated less viral integrations at TSSs and CpG islands (Table 2). To investigate further, viruses extracted from this tumor were used to infect 293mCAT cells and NGS integration site analysis was performed after 14 days. The MLV IN TP$^-$16 integration profile in 293mCAT cells paralleled that seen during viral spread following plasmid transfection with MLV IN-XN (Table 3) (Fig17) Cumulatively, this indicates that the viral population from the TP$^-$16 tumor maintained the MLV IN TP$^-$ genotype as it

Table 3. Comparison of integration sites in 293mCAT cells of MLV IN XN and MLV IN TP$^-$ derived from TP$^-$16 mouse viruses.

	Unique sites	TSS+/-1kb (%)	CpG+/- 1kb (%)
WT[a]	64828	14208 (21.9)	18864 (29.1)
IN XN[a]	37638	2029 (5.4)	3171 (8.4)
IN TP$^-$16[b]	47093	2959 (6.28)	4541(9.6)
MRC	10000	169 (1.69)	270 (2.7)

[a] Viral spread through the plasmid transfection

[b] Viral spread by infection of viruses extracted from mice

spread within the mouse. A more detailed comparison of integration site preference of the WT6 and TP^{-}16 tumors, which share a DoD of 34 days, is presented in Fig 17. The WT MLV integrants are symmetric around the TSS (black, Fig 17A left), in contrast to the non-infected control (NC, orange Fig 17A right). While the TP^{-}16 integrants are concentrated around the TSS, this distribution is asymmetric and more dispersed than in the WT6 tumor (Fig 17A blue, center) with secondary integration peaks +/-8 kb from TSSs. Integration profile of MLV IN-XN viral spread via plasmid transfection was also compared to the integration profile of 293mCAT infection of viruses derived from TP^{-}16 tumor (Fig17 B). Both profile show the same percent integrants at TSS and symmetric distribution. Some limited ChIP-Seq data is available for mouse BET proteins, and integration of WT6 and TP^{-}16 virus at Brd4 binding sites was considered (data taken from ENCODE ID GSM1262345, murine AML MLL-AF9/NrasG12D cells) (Fig 18A). Interestingly, an approximately 20% decrease of MLV IN TP^{-}16 sites was observed at Brd4 sites, as compared to MLV WT6 (Fig 18A). Furthermore, approximately 50% less TP-16 integrations were observed +/- 1 kb from the histone modifications H3K4me1 and H3K4me3 when compared to WT6 (ENCODE IDs ENCSR000CCI and ENCSR000CCJ, respectively), and were at a level that was comparable with the non-infected control (NC) (Fig 18B). H3K4me1 and H3K4me3 are considered MLV supermarkers (164), with H3K4me3 being associated with nucleosome-bound BET proteins (165).

In addition to the histone marks studied above, BET proteins are highly associated with active enhancer features, specifically acetylated histone tails. In the human lymphoma cell line Ly1 DLBCL, 79.1% of H3K27Ac sites overlap with Brd4 and 92.2% of chromatin bound Brd4 sites are at regions with the H3K27Ac active enhancer histone mark (166). MLV integrations are reported to be highly enriched at H3K27Ac sites (69, 76, 154, 167). Analysis of the overlap between MLV WT6 SITES with H3K27Ac peaks

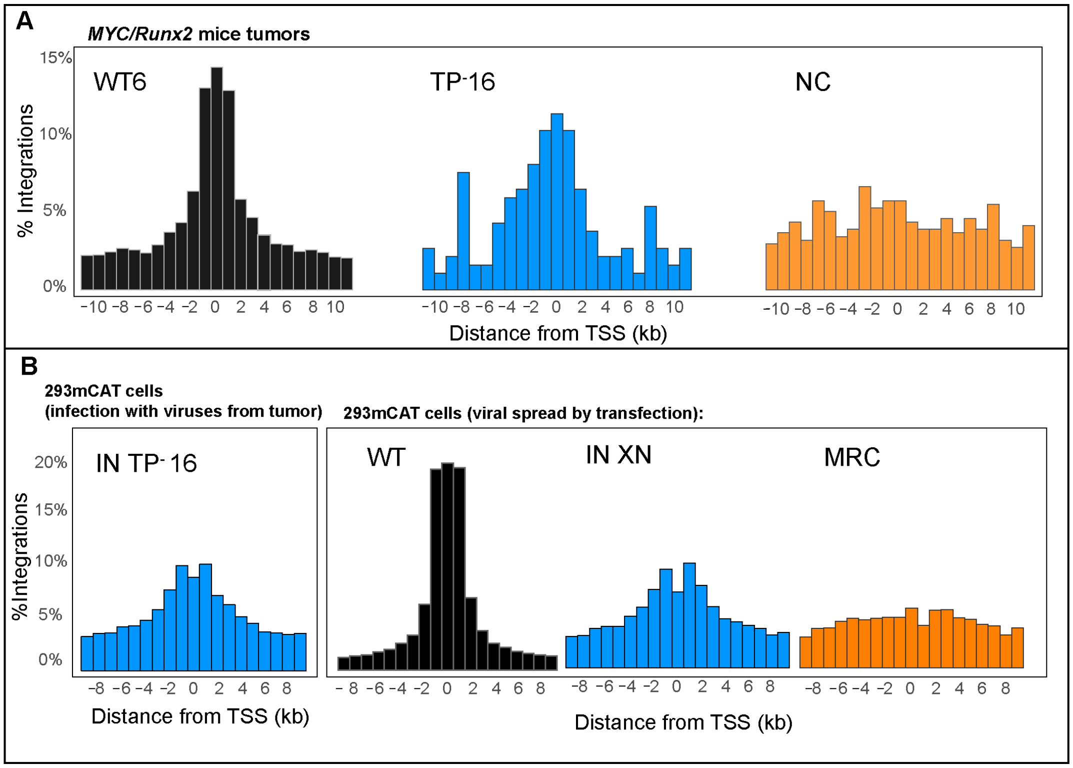


Fig 17. Comparison of MLV integration profiles of IN TP^{-}16 with WT6 mouse and 293mCAT cells infected with WT MLV. In all panels, the IN TP^{-}16 integrants are indicated in blue, WT6 in black/grey, and non-infected control (NC) in orange. (A) Histograms of MLV integration profile of tumors at the nearest TSSs. Percentage of RISs plotted against annotated TSSs compared to NC. (B) Histograms of MLV integration profile of 293mCAT cells infected with IN TP^{-}16 tumor derived viruses and 293mCAT infected with WT MLV and MLV IN XN.

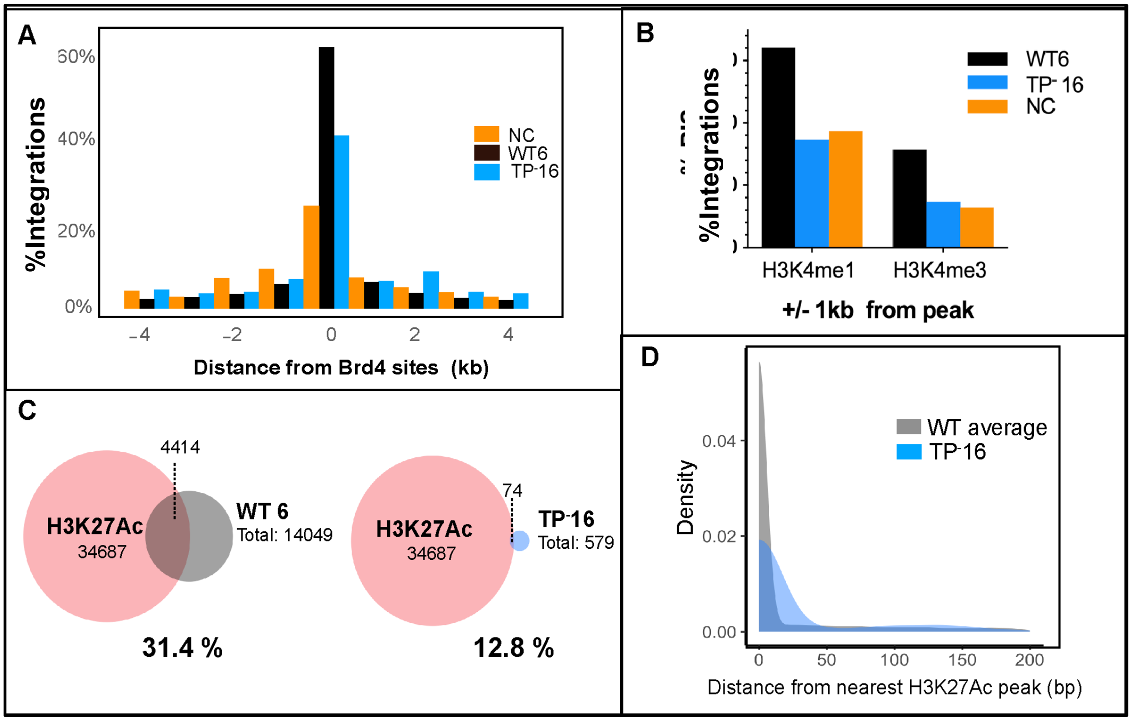


Fig 18. Comparison of TP^{-} 16integrations sites overlap with Brd4, H3K4me1/3 and H3K27Ac ChIP-seq peaks of IN TP^{-}16 with WT6 mouse. In all panels, the IN TP^{-}16 integrants are indicated in blue, WT6 in black/grey, and non-infected control (NC) in orange. (A) Percentage of RISs plotted against annotated TSSs compared to NC. Association of integration sites with Brd4 binding regions. Percentage of RISs plotted against annotated Brd4 binding sites from GSM1262345. (B) Percent integration sites overlap within +/- 1 kb from H3k4me1/3 peaks. (C) Venn diagram of overlap of H3K27Ac peaks (ENCFF974HMO) with RISs. The dash lines indicate the number of RISs overlapping with H3K27Ac. (D) Histogram of RISs from the nearest H3K27Ac peaks. The combined WT RISs from all tumor samples (grey) is plotted against TP^{-}16 (blue).

(ENCODE ID ENCFF001KYG) (Fig 18C) showed that 31.4% of integrants were at H3K27Ac sites, although this was markedly reduced with MLV IN TP$^-$16 RISs, which exhibited 12.8% overlap. Furthermore, the distribution of the distance between RISs and the nearest H3K27Ac peak was significantly broader for TP$^-$16 than for the cumulative RISs from all WT tumors (Wilcoxon Rank Sum, p<2.2e-16; Fig 18D). All of the WT samples had similar percent overlaps (31.2% for WT8, 35.3% for WT10, 37.1% for WT12) (Table 4). IN TP$^-$16 has significantly less percent overlap from the WT average according to 1 sample t-test (p = 0.000367). Overall, these tumor integration profiles indicate that TP$^-$16 RISs are generally located further away from promoter and active enhancers than those of WT MLV.

Viruses extracted from MLV IN TP$^-$16 tumor were used to infect 293mCAT cells and integration site analysis was performed after 14 days. The integration profile in 293mCATs for MLV IN TP$^-$16 paralleled that of MLV IN-XN in tissue culture (Table 3, Fig 17B) therefore the same phenotype with respect from TSSs. Overall, these integration profiles from the tumors indicate that the viral population from the TP$^-$16 tumor maintained the MLV IN TP$^-$ genotype within the mouse and that the integrants are distributed further away from promoter and active enhancers compared to the WT MLV.

The top copy number RISs from MLV IN TP$^-$ target known MLV integration CIS

Table 5 lists the top six copy number RISs present in the IN TP$^-$16 tumors, which represent the predominant integrants during clonal expansion in the tumor (90). Significantly, four of the top six are in three WT MLV common integration sites identified previously in the MYC/Runx2 mouse (*Mapk13, Ccnd1, Hdac6)* (90). *Ccnd1* is significantly upregulated in this tumor model compared to normal thymus, and *Rasgrp1* and *Hdac6* are two of the most frequent CIS targets previously identified (90). The genomic positions of the TP$^-$16 RISs from this study at these three genes in the mouse

Table 4. Percent overlap of RISs from mouse tumor and H3K27Ac pea

Samples	Unique sites	Total H3K27Ac peaks	Number of RISs overlapping with H3K27Ac peaks	% RISs overlapping with H3K27Ac peaks
WT6	14049	34687	4414	31.4
WT8	4319	34687	1348	31.2
WT10	5043	34687	1779	35.3
WT12	1422	34687	528	37.1
$TP^{-}4$	3366	34687	1125	33.4
$TP^{-}6$	2370	34687	724	30.5
$TP^{-}7$	7544	34687	1808	24.0
$TP^{-}9$	1601	34687	427	26.7
$TP^{-}16$	579	34687	74	12.8

Table 5. Top 6 copy number RISs of MLV IN TP^{-}16 tumor

	Strand(-/+)	Integrant position			Copy number
*_Hdac6_ (exon 2)	-	chrX	exon	7946933	62,234
*_Hdac6_ (intron)	-	chrX	intron	7930719	23,787
Mapk13	+	chr17	intron	28773031	54,637
*_Rasgrp1_	-	chr2	intergenic	117,434,635	12,930
*_Ccnd1_	-	chr7	Intergenic	144940638	19,122
Gng7	-	chr10	intron	80957684	16,963

* MLV CIS genes

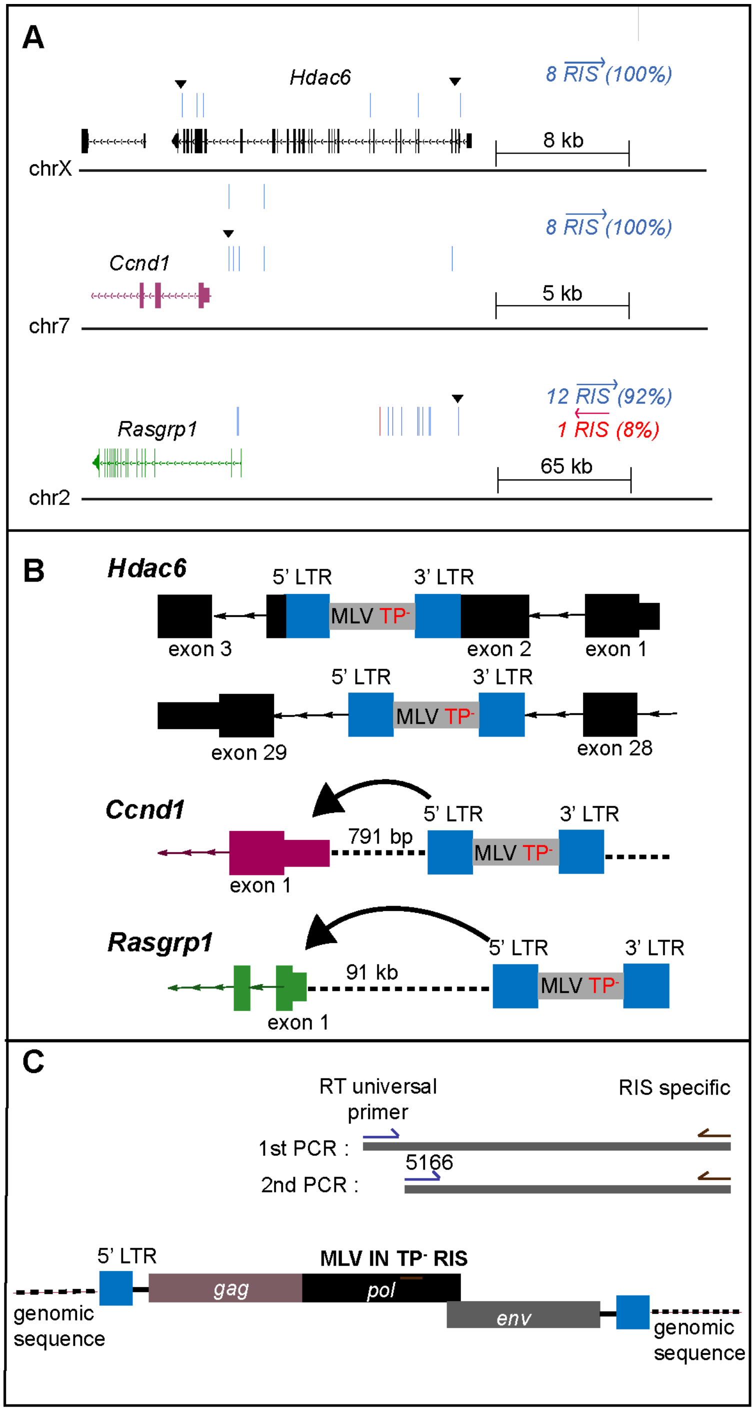


Fig 19. Analysis of the top six copy number RIS within the IN TP$^-$16 tumor. (A) Orientation bias of RIS in three CIS genes: *Hdac6* (black), *Ccnd1* (purple), *Rasgrp1* (green). The exons and introns are represented in boxes and lines with arrow that show strand orientation respectively. RIS is represented by a vertical bar and differentially colored based on orientation (blue forward, red reverse) relative to the plus-strand DNA. Gene structures are derived from Integrated Genome Browser (mm10). (▼) denotes the top copy number RIS (B) Diagram

of the TP$^-$16 RISs relative to the closest CIS gene. The integrated MLV is depicted in grey, with LTR indicated in blue. Coding regions of CIS are represented as in panel A. Intergenic regions are represented in dashed lines. Direction and distance between 5' LTR and TSS is indicated. IN genes verified to maintain the TP$^-$ phenotype are indicated. (C) Schematic diagram of the nested PCR utilized to isolate the RIS from mouse tumor DNA. Primers used in the first and second round of PCR are included in the diagram. RIS specific primers were designed based on mouse genome (mm10). Sequences of all oligonucleotides are described in Table 1.

is shown schematically in Figs. 19A and 19B. Remarkably, insertions at all three genes show an orientation bias towards the sense direction. For *Ccnd1* and *Rasgrp1*, this orientation bias is antisense to the host gene transcription, which is consistent with enhancer insertional activation reported previously (93, 95, 168) Interestingly, the *Rasgrp1* cluster of integrations map >91 kb upstream of the *Rasgrp1* promoter, whereas for *Ccnd1*, the high copy number insert is located 791 bp from the promoter (black triangle, Fig 19A). It is striking that two independent insertions in *Hdac6* were highly abundant, and both were intragenic; the first disrupted exon 2 at the N- terminus of the protein, while the second was located within the intron spanning exons 28 and 29.

It was of considerable interest to verify that the MLV inserted at these three genes maintained the IN TP$^-$ genotype. Based on the known insertion site, a nested PCR was developed to allow amplification of the IN gene through the specific host junction (Fig 19C and Table 3). Sequencing of the resulting PCR products verified that all four insertions at *Hdac6* (2 insertions), *Ccnd1*, and *Rasgrp1* maintained the parental IN TP$^-$ mutation. Interestingly, when the 3' LTR junctions of 3 of the top 6 RISs (*Mapk13, Ccnd1,* and *Hdac6)* were sequenced, heterogeneity was observed at the positions +1/+2 downstream of the MLV junction sequence (TCTTTCA). Viral integration involves the cleavage of the LTR terminal dinucleotide, exposing the conserved 3' CA region and generating a 5' single stranded (ss) tail on the viral DNA substrate. The observed sequence heterogeneity corresponded with either the predicted host DNA sequence (Fig 20A, black text in parenthesis) or the sequence encoded through the 5' ss viral tail (Fig 20A, blue in parenthesis). Mixed populations at this position have not been previously identified, and may reflect unique repair mechanisms at the viral/host DNA junction in the *MYC/Runx2* mice model, in which p53 activity is reported to be suppressed (Fig 20B) (99, 100).

Discussion

Recombination with the endogenous polytropic *Pmv20* virus was detected in multiple tumors (IN TP$^-$4, 6, 7, 9, 15, 18, and 19). While this recombination restored the IN TP, the presence of recombinant virus did not correlate with the mouse DoDs; IN TP$^-$16 demonstrated no recombination with polytropic virus (DoD 34 days), whereas IN TP$^-$4, 6, 7, 9, 15, 18, and 19, which did undergo recombination, had DoDs of 50, 57, 37, 63, 33 and 36 days, respectively. Our analysis cannot determine the point in the infection time course at which recombination occurred, and it is possible that the recombination event for IN TP$^-$9, for example, which had a DoD of 63 days, occurred late in tumor development and thus did not have a marked effect on tumor progression. It is interesting to note that following infection with IN TP$^-$, no mice survived beyond day 70, whereas the uninfected control mice survived beyond 110 days, suggesting that the presence of a recombinant virus may have affected long-term survival.

In the mouse model, we cannot eliminate the possibility that the integration at active promoter/enhancer regions observed in IN TP$^-$ is due to transcomplementation by endogenous Gag-Pol proteins from the xenotropic MLV *Bxv-1* locus (101). C57BL mice have been documented to express xenotropic MLV at low levels in vivo, which can be induced in tissue culture with IdU (113, 169). Although the xenotropic Env would exclude infection of these endogenous viruses, complementation of the IN protein *in trans* cannot be eliminated. This could result in the low level of bias towards the TSS observed within the *MYC/Runx2* tumors, in the absence of recombination events. This issue of transcomplementation by xenotropic MLV would potentially not be a problem in the use of IN TP$^-$ virus in non-murine cells. However, other species do have their own ERVs, which may or may not contribute to the target-site selection of MLV based vectors.

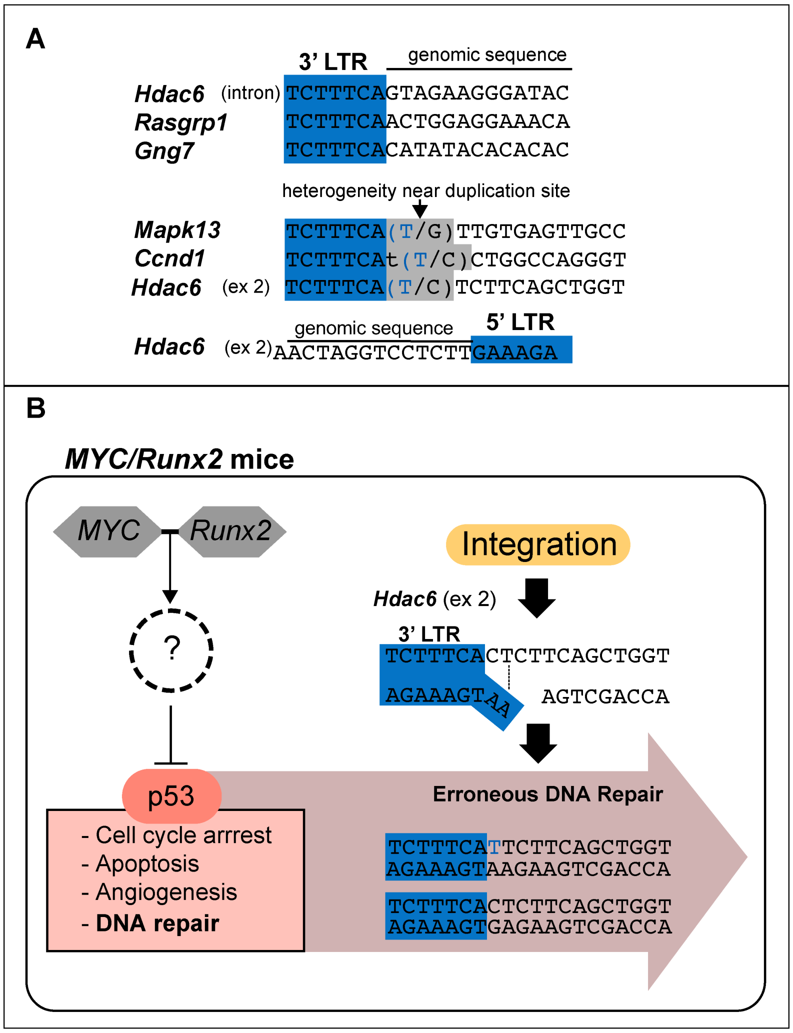


Fig 20. **Base heterogeneity near the LTR junction.** (A) Sequence of the LTR/mouse genome junctions of six TP^{-}16 RIS. The MLV LTR termini are shaded in blue and the mouse genomic sequences are unshaded. Bases shaded in grey displayed a mixed population using Sanger sequencing; bases corresponding to the viral LTR are in blue and those of the mouse genome are in black. Base in lowercase can be both from LTR and mouse genome. (B) Proposed influence of the transgenes *MYC* and *Runx2* on genetic and phenotypic landscape in mice. The 3' LTR junction of *Hdac6* (exon 2) integrant is shown. Expression of both *MYC* and *Runx2* transgenes synergistically suppress p53 activation through unknown mechanism of activation of other genes that neutralizes p53 function. p53 functions as a regulator of many cellular processes such as cell cycle arrest, apoptosis, angiogenesis and DNA repair. Suppression of p53 could affect DNA repair mechanisms and cause erroneous DNA repair on the provirus.

Table 6. Copy number of the top TP^{-}16 RISs in *Hdac6, Ccnd1, Mapk13.*

CIS gene	Integrant Identifier*	Copy Number	Central base pair position
Hdac6 (exon)	M00851:172:000000000-B3VWL:1:2105:17280:1567	62234	7946931-7946932
Hdac6 (exon)	M00851:172:000000000-B3VWL:1:2103:15122:23186	1	7946930-7946931
Ccnd1	M00851:172:000000000-B3VWL:1:2105:9498:5837	1	144940635-144940636
Ccnd1	M00851:172:000000000-B3VWL:1:2105:15456:1380	19122	144940636-144940637
Mapk13	M00851:172:000000000-B3VWL:1:2105:27325:19943	4	28773027-28773028
Mapk13	M00851:172:000000000-B3VWL:1:2105:13737:1495	54637	28773029-28773030

*Representative sequence identifier of the integrant is shown when multiple copies are present.

Another source of recombination with ERVs that wasn't analyzed is the ecotropic ERV present in the mice. C57BL/6 has one endogenous ecotropic MLV, *Emv2* that is poorly expressed and has a replication defective RT (101). It can be rescued by recombination with replication competent MLVs (170, 171).

Analysis of the top retroviral integration sites (RISs) copy number sites in TP$^-$16 tumors that maintained the IN TP mutation revealed anomalies in target-site duplication. The high frequency (3/6) of a heterogeneous mixture of two bases within the target-site duplication region adjacent to the U5 LTR is of interest. In two cases, the sequence read a mixture of C and T, where the host target DNA encoded a C. The third case involved a mixed G/T population, where the host DNA encoded a G. In all three cases, the heterogeneity was observed within two bases of the viral CA terminus. The *MYC/Runx*2 mouse model is hypothesized to function to overcome p53 (90) (Fig 20B). p53 is known to interact with components of the nucleotide excision repair, base excision repair (BER), mismatch repair (MMR), and homologous recombination (HR)/NHEJ pathways (172), and repair of the target DNA site occurs using host repair mechanisms. It is possible that the observed heterogeneities have not been previously identified under conditions of low p53 activity. The incorporation of a T at these positions could be the result of base-pairing of the 5'-tail of the viral genome, encoded by TT for MLV. If these mismatched viral bases are ligated to the 3' extended gap-repair product, rather than undergoing strand-displacement and excision, the observed heterogeneities could be maintained in the tumor population after DNA replication. The G-to-T transversion observed at the *Mapk13* integrant could be a result of base-excision repair (BER), if integration was directed into a region of oxidative damage, for example at an 8-oxo-dG (173). In HIV, oxidative BER proteins directly influence integration at the sequence level.(173).

Heterogeneity at the target duplication sites was observed for integrants at the *Hdac6*, *Ccnd1*, and *Mapk13* loci. Although the mapping studies indicated that all three of these loci had a secondary integrant 1-2 bases away from the first (Table 6), only one integrant spanning the viral:host junction was amplified. Analysis of the integrant copy number within the NGS libraries indicated a much higher incidence of one of integration sites at each locus, and it was the predominant integration in the population that was amplified and analyzed. The heterogeneity at the target site duplication was therefore not linked to the presence of a neighboring integrant in the population.

For retroviruses, enhancer activation usually occurs upstream of the gene in the antisense orientation, or downstream in the sense orientation (168). Indeed, this was the orientation bias observed for both *Ccnd1* and *Rasgrp1*. For *Ccnd1,* the top integrant is included in the integration cluster near the promoter (Fig 19A). Similarly, previous studies of WT MLV in *MYC/Runx2* tumors indicated a cluster of insertions at the 5' end of the *Ccnd1* gene, predominantly upstream of the coding sequence (90). The proximal cluster initiates overexpression via retroviral enhancer elements (174). *Ccnd1* has an important role in cell cycle regulation, and overexpression induces the formation of different cancer types (175-177). For *Rasgrp1*, the integrant analyzed was 91,758 bases upstream of the *Rasgrp1* promoter, and oriented with the viral promoter in the opposite orientation. Activation of *Rasgrp1* most likely occurs through an enhancer activation event. Sequence analysis indicated this integrant maintained the IN TP$^-$ sequence. It was initially surprising to see the RIS orientation bias >91 kb upstream of *Rasgrp1* promoter. For human leukemia virus -1 (HTLV-1), long-range interactions between target gene promoters and viral enhancers are facilitated through chromatin looping utilizing the host zinc finger binding protein, CTCF (178, 179). The HTLV-1 provirus contains a CTCF nucleotide-binding motif that has been shown to mediate clone-specific deregulation of host transcription from distances up to 300 kb from the provirus (179), and CTCF-mediated *cis* contacts within the host genome can be as far as 1.4 MB (178, 179). Although M-MLV does not encode a known CTCF binding motif, CTCF binding sites have been identified at the promoter region of *Rasgrp1* and ~10 kB downstream from the RIS (ENCODE reference ENCFF310MUQ). CTCF-mediated transcription varies depending on the cell type. Validation of CTCF binding would require circular chromosome conformation capture (3C or ChIA-PET) analysis from the tumors, which is not available for this study. However, the presence of these CTCF binding sites provides a potential mechanism for the MLV viral enhancer to interact with *Rasgrp1* promoters that are distant from each other, thereby driving overexpression of *Rasgrp1* concomitant with tumorigenesis in these mice (180-182). For *Ccnd1*, CTCF-dependent long-range loops have been identified that reposition distal clusters of retroviral insertions, driving gene activation (174).

The *Hdac6* gene also displayed a biased integration in an orientation opposite that of transcription, however these integrants are within the *Hdac6* gene. For IN TP^{-}16, the two most abundant integrants in the library map within exon 3 at the 5' end and within an exon at the 3' end of the gene. The results imply a loss of function through oncogenic selection, however the mechanism cannot be determined. *Hdac6* is reported to interact with *Runx2* (183, 184) as well as being involved in multiple cellular processes, including organization of the immune synapse, cell migration, protein degradation, and viral infection (185).

Section III. Identification of chromatin state preference of MLV IN TP- integrations in K562 cells

MLV IN TP$^-$ integration sites at chromatin states and histone modifications in K562 cells

To further investigate MLV IN TP$^-$ integration, a single round of infection of human leukemia cell line K562 with this virus was performed. These cells recapitulated the decreased integration percentage at TSSs and CpG sites that was previously observed in 293mCAT cells for the IN-XN construct (Table 3 and Table 7, Fig 21). To extend this analysis, 15 chromatin states and histone modifications were considered. The use of chromatin states provides a different approach to understand the genomic landscape, as previously reported (75). In this approach, clusters of chromatin marks are used to define functionally active states of chromosome, which are specific for each cell line. For K562, 15 chromosome states have been utilized to analyze MLV integration (75), and 85% of integrants were shown to map to strong enhancers and active promoter regions. The overlap between MLV WT and TP$^-$ RIS and the components of the 15-chromatin state model was investigated in K562 cells. As shown in Fig 22A, 73.8% of MLV WT integrations mapped to the same three highest states that were identified previously (75), which were annotated as active promoter (state 1) and strong enhancer (states 4 and 5). In contrast, the MLV IN TP$^-$ integrations displayed a divergent integration preference, with the top 2 chromatin states being heterochromatin (State 13; 21.6%) and weakly transcribed regions (State 11, 19.4%) (Fig 22B, bottom). Although not amongst the top states, the loss of the IN TP did not eliminate integrations at active promoters (State 1) and enhancers (Strong enhancers, states 4 and 5; weak enhancers, state 7), which cumulatively accounted for 35% of the IN TP$^-$ integrants.

Table 7. MLV integration site mapping of single round infected WT MLV and MLV IN TP$^-$ in K562 cells

Sample	Unique sites	TSS+/-1kb (%)	CpG+/- 1kb (%)
WT	4384	1160 (26.46)	1417 (32.32)
IN TP$^-$	934	68 (7.28)	98 (10.49)
RIC*	10000	169 (1.69)	270 (2.70)

*RIC – Random integration control

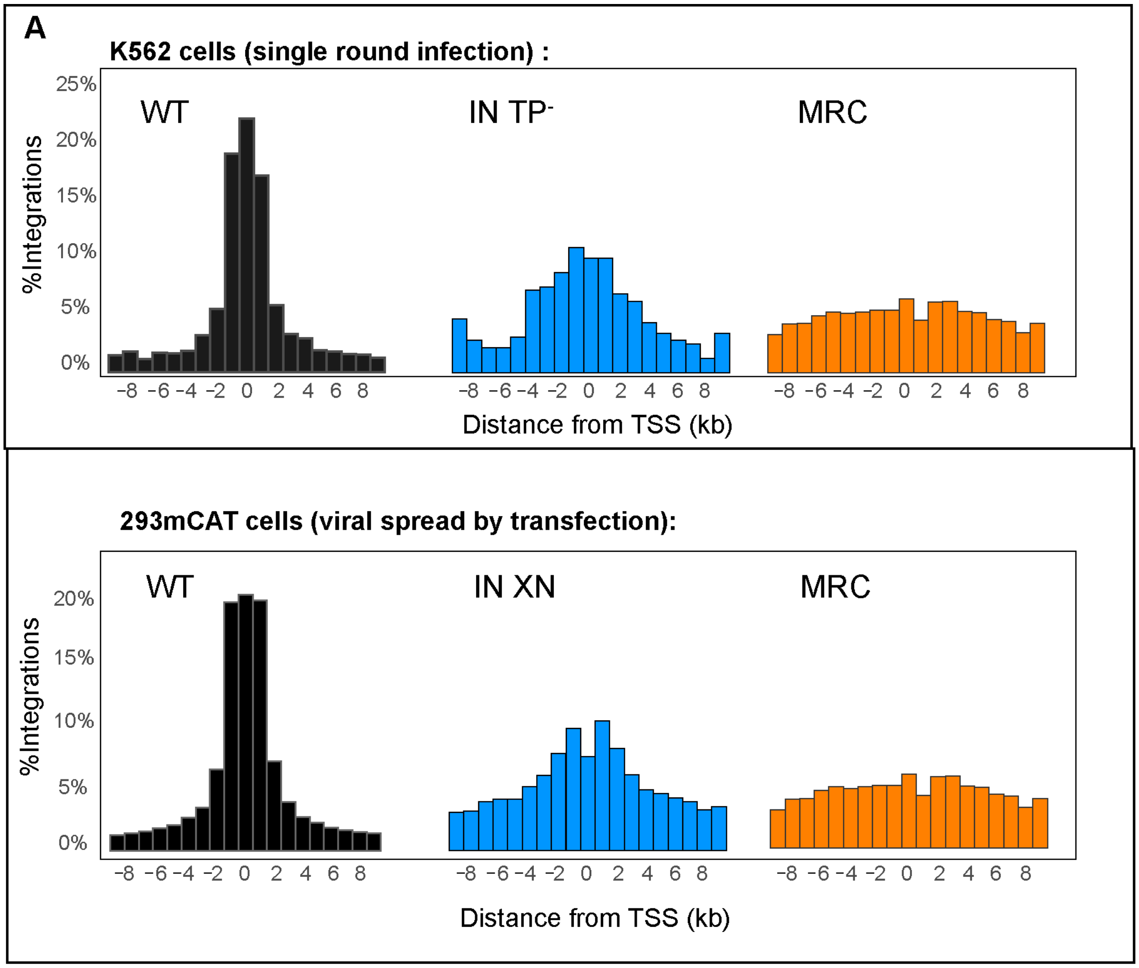


Fig 21. Histogram of integration in K562 cells infected with MLV IN TP^{-} and WT MLV from TSSs. In all panels, the IN TP^{-}16 integrants are indicated in blue, WT6 in black, and RIC in orange. (A) Histograms of MLV integration profile at the nearest TSSs. Percentage of RISs plotted against annotated TSSs compared to RIC. (B) Histograms of MLV integration profile of WT MLV and MLV IN XN in 293mCAT cells at the nearest TSSs.

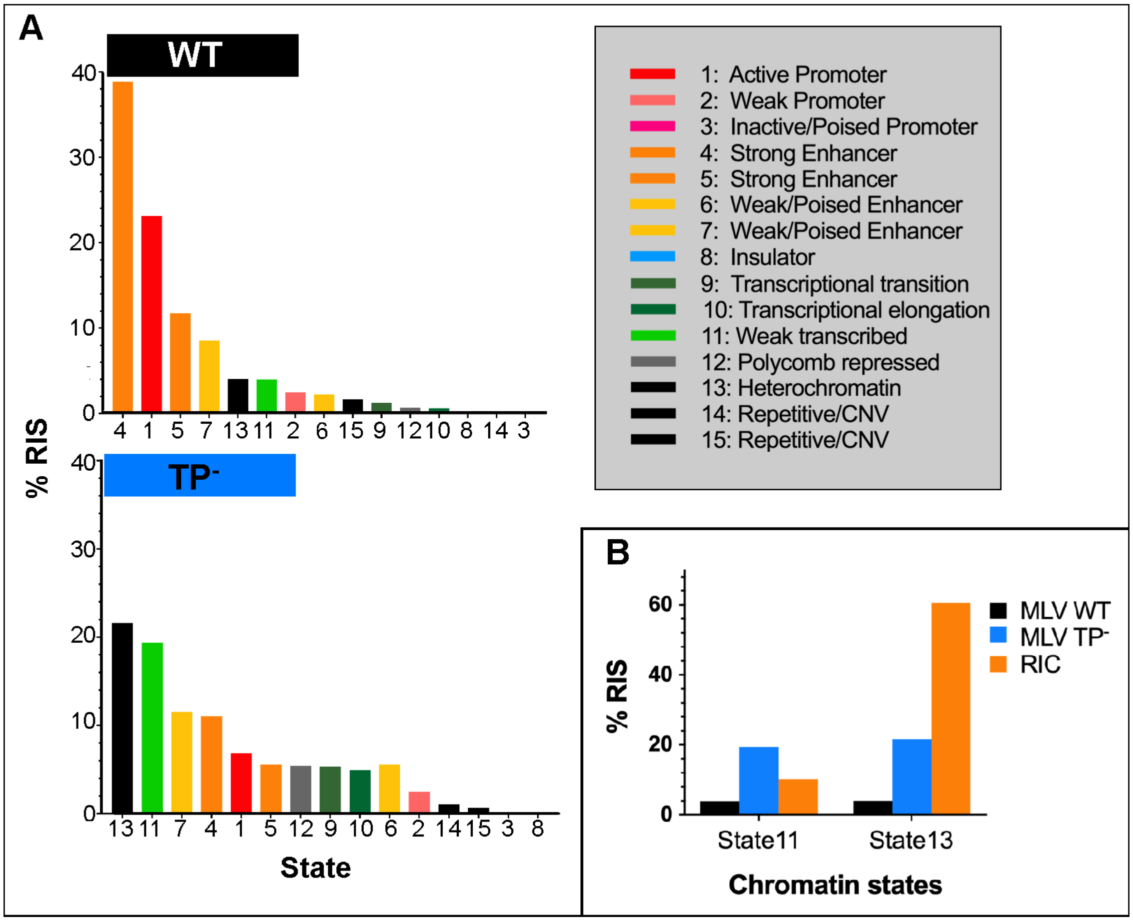
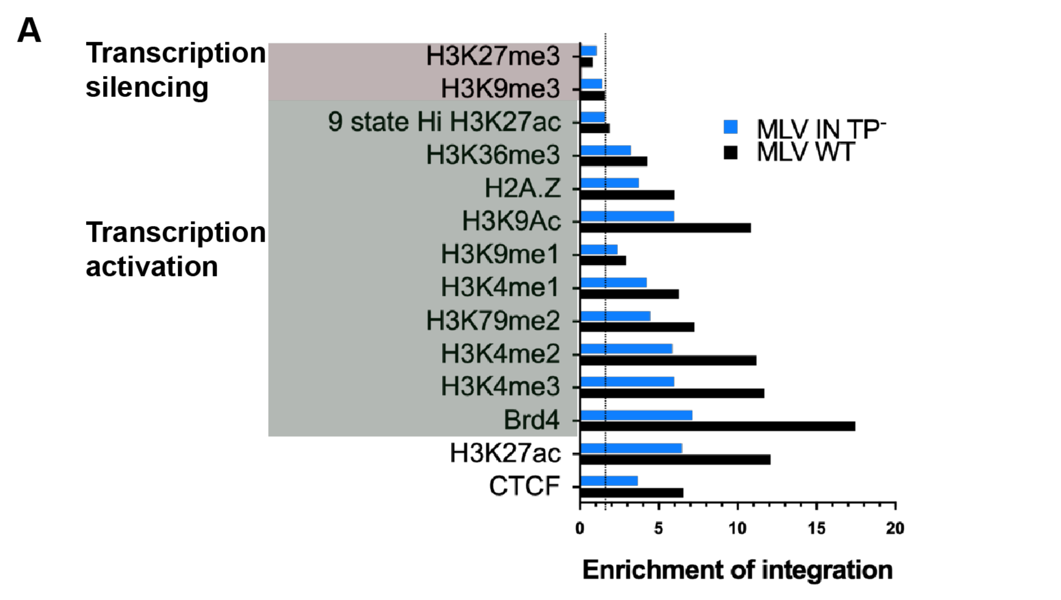


Fig 22. Chromatin profile of MLV IN TP$^-$ integrations in K562 cells based on the 15-state chromatin model

In all panels, the IN TP$^-$16 integrants are indicated in blue, WT6 in black, and RIC in orange. (A) Percentage of RISs in 15-chromatin states (151) in K562 cells: WT infection (top panel) and MLV IN TP$^-$ infection (bottom panel). Each chromatin state is labeled with corresponding color as indicated. (B) Percentage of RISs in state 11 and state 13 compared to RIC. State 11 is described as weakly transcribed regions and state 13 are heterochromatin regions (151).

The median chromosomal coverage in K562 for chromatin states 11 and 13 were reported to be 11.3 and 71.4% (151), respectively, which correlates well with the genome coverage of computer-generated integrations within the random integration control (RIC; Fig 22B). For state 11, corresponding to weakly transcribed regions, loss of IN TP increased integration frequency 4.8-fold compared to WT (Fig 22C; 19.4% for IN TP$^-$; 4.01% for WT MLV). Significantly, this frequency is 2-fold above the RIC, indicating a bias for integration into these weakly transcribed regions. Similarly, integration into heterochromatin (state 13) increased ~ 5-fold in MLV IN TP$^-$ (21.63%) compared to WT MLV (4.06%). For comparison, WT MLV integrations at heterochromatin are lower than RIC. Thus, in the absence of the IN TP, MLV integration preference for heterochromatin and weakly transcribed regions increases, while integration at active promoters and enhancers was disfavored.

The chromosome states described above are defined, in part, through profiling combinations of a set of histone modifications modifications and the occupancy of various cis-regulatory elements by known protein factors (151). Directed by the bromodomains, acetylated histone modifications are important in determining BET proteins interactions; however, additional marks including H3K4me2/3 are elevated in Brd-bound nucleosomes (165). MLV integrations have been strongly associated with H3K4me1, H3K4me2, H3K4me3, H3K27ac, H2Az and H3K9ac modified chromatin (69, 70, 75, 76, 163, 167). Fig 22 show the proportion of WT and IN TP$^-$ integrants found in regions marked by the various epigenetic marks and proteins associated with integration site preference in K562 cells. ChIP-seq data for all chromatin modifications was obtained from the ENCODE consortium, and enrichment of viral integration over RIC control was computed for each modification. Significantly, the largest decrease in enrichment on loss of the IN TP region was associated with Brd4 binding sites (17.5-fold for WT vs 7.12-fold



B

	% Integrations			Fold enrichment	
	WT	IN TP-	RIC	WT	IN TP-
Brd4	81.4	33.2	4.7	17.5	7.1
CTCF	53.4	29.8	8.2	6.5	3.6
H2A.Z, enhancers/promoters	91.2	56.6	15.2	6.0	3.7
H3K27ac, active reg elements	89.8	48.1	7.4	12.1	6.5
H3K9Ac, active reg elements	89.9	49.6	8.3	10.8	6.0
H3K36me3, active gene bodies	71.7	54.7	16.8	4.3	3.3
H3K79me2, active gene bodies	71.8	44.0	9.9	7.3	4.5
H3K4me1, enhancers	92.3	61.9	14.7	6.3	4.2
H3K4me2, promoters/enhancers	90.3	47.3	8.1	11.2	5.9
H3K4me3, promoters	88.8	45.4	7.6	11.7	6.0
H3K9me1, activation	77.7	63.9	26.5	2.9	2.4
H3K9me3, repressed regions	49.8	42.9	30.9	1.6	1.4
H3K27me3, repressed regions	17.6	23.3	22.0	0.8	1.1

Fig 23. Enrichment analysis at different histone modifications of MLV IN TP$^-$ integrations in K562 cells In all panels, the IN TP$^-$16 integrants are indicated in blue and WT6 in black. (A) Enrichment of integrations in ChIPSeq peaks of different histone marks and Brd4 binding regions in K562 cells. Value of enrichment is calculated by dividing the number of RISs with the number of RIC at each histone mark. The dotted line is the level of enrichment expected by chance. Transcription silencing histone marks are in pink and transcription. (B) Comparison of percentage of integrations for WT, IN TP$^-$ and RIC within the ChIP-seq peaks for different histone modifications, CTCF and Brd4.

for MLV IN TP^{-}) (Fig 23). The WT virus was most highly enriched at the epigenetic modifications H3K4me3, H3K4me2, H3K27ac and H3K9ac, which in consistent with previous report (75). Interestingly, loss of IN TP resulted in a >45% decrease in fold enrichment at H3K4me3, H3K4me2, and H3K27ac sites, which is consistent with a loss of association with BET proteins.

Discussion

The absence of IN TP reduced the integration bias towards strong enhancers and active promoters and was preferentially associated with heterochromatin and weakly transcribed regions. The secondary preference towards active regulatory elements (state 1, 4 and 5) is maintained, which corresponds to median genome coverage of only 2.5%. This observed integration of MLV IN TP^{-} at active promoter/enhancers could be the result of a second IN site interacting with BET proteins (154).

The heterochromatin and weakly transcribed regions in the 15-state model share the absence of the chromatin marks H3K4me1/2/3, H3K27me3, H3K27ac, H3K9ac and CTCF, with the weakly transcribed state 11 containing low levels of H3K36me3, H4K20me1 (151). A more defined model with 25 and 50 chromatin states has recently been described, which makes further subdivisions based on an expanded set of chromatin marks. In these, the heterochromatin state is distinguished from the quiescent state by the presence of the H3K9me3 mark. While K562 ENCODE data is not available for the full-expanded set of chromatin marks, we observed no significant fold enrichment of WT MLV or IN TP^{-} integrations at H3K9me3. Therefore, most of the MLV IN TP^{-} integrations in heterochromatin would reasonably be categorized as targeting the quiescent chromatin state in the 25- and 50- chromatin state model. The quiescent state is defined by the large absence of any histone modifications, similar to the heterochromatin state in 15-chromatin state, an inactive state with low annotated non-coding and coding transcripts (186).

Although many studies have documented the role of the IN TP in driving integration towards active promoters and enhancers (68, 75, 76). this is the first study to define where the integrations are directed in the absence of the BET protein interaction. Two models could explain the integration of 40% of MLV TP$^-$ integrants into regions with limited histone modifications. Firstly, the MLV IN may display an innate recognition of unmodified histone tails, or secondly, modified histones may present a steric hindrance for IN binding. Both MLV and prototype foamy virus (PFV) encode an N-terminal extension domain (NED) (61, 77), however PFV IN does not encode a homologous TP. The PFV IN CCD-CCD interface interacts with the H2A-H2B heterodimer, specifically with the C-terminal helix of H2B and N-terminus of H2A. Three PFV IN residues (P135, P239 and T240) were shown structurally to interact with H2B, and they reside in the loops between β-sheet 1/2 and α helix 4/5 (64, 187). Binding to the nucleosome results in a 7Å deformation of the target DNA, and ultimately drives viral integration into heterochromatin regions, Lamin A/B1 rich-regions, and intergenic regions (64, 187). H2A and H2B are the most diverse histones, which, along with distinct post-translational patterns, contribute to the complexity and variability of H2A-H2B dimers (188). It is possible that the MLV IN CCD-CCD dimer may interact with a specific variant of H2A-H2B heterodimer with distinct modifications independent of BET proteins.

Transposing the whole α helix 4/5-loop region of the PFV IN CCD (RPTK) into MLV IN resulted in viral titer equivalent to those of the catalytically inactive MLV D184N (data not shown), however, point mutations (S239P and R240T) and β-sheet 1/2 loop exchange (GLY → PSQ) produced viral titers. This implies that the homologous loop region of MLV IN CCD (GSRD) is integral to IN stability or is a site of secondary interactions. When Asp from MLV IN CCD α helix 4/5-loop was placed instead of Lys of

PFV (GSRD → RPTD), it surprisingly resulted in viral titer. The integration profile of viral infection in 293mCAT cells for both the MLV IN CCD (GLY→PSQ) and MLV IN CCD (GSRD→RPTD) were analyzed (data not shown). Both showed decreased integration at TSSs and CpG islands, but not at the same level as MLV IN-XN, suggesting the BET interaction with WT IN TP has a stronger influence on the target-site selection. When these mutations and CCD loops were introduced in MLV IN-XN, however, only the point mutations produced viral titer and can be used for future integration profile studies.

Summary, Conclusions and Future Directions

Multiple studies have explored the MLV IN protein interaction with host BET proteins and how it directs integration into sites of active expression marked by acetylated chromatin. This led to studies of the effect of loss of MLV BET binding site on the pathogenesis of the virus. In this study, we infected *MYC/Runx2* mice, a transgenic model displaying rapid tumor formation that is further accelerated by MLV infection with MLV lacking the IN tail peptide. Globally, the infection time course for the IN TP^{-} cohort showed a biphasic DoD curve, with 1/3 of the mice developing tumors early, as with WT infection, and 2/3 of the mice developing tumors later, paralleling the uninfected *MYC/Runx2* control mice. All the mice died by day 73 compared to the control mice (day 115). Median survival of IN TP- infected mice (50 days) is longer than WT infection (34 days) with We extensively characterized the mouse IN TP^{-}16, which displayed early tumorigenesis and showed no detectable signs of recombination with endogenous viruses. Characterization of TP^{-}16 tumor viral integration sites indicated a broader integration profile around TSS and decreased association with H3K27ac, H3K9me1 and H3K9me3 marks (Figs 18). Characterization of the virus at specific integration sites of high abundance indicated that the IN mutation was preserved, and the characteristic target-site profile for IN lacking the TP was maintained following 293mCAT cell infection with IN TP^{-}16 tumor-derived virus. However, integrations into CIS known to accelerate tumor formation were present, suggesting that stochastic integration into an open chromatin hotspot can still provide the positive selection required for tumor outgrowth.

The overall goal of these experiments was to determine whether MLV IN TP^{-} virus represent safer gene delivery vectors, using an *in vivo* mouse model. The uniform acceleration of tumorigenesis by MLV infection in the *MYC/Runx2* mouse was not observed with IN lacking the BET interaction domain at its C-terminus. A major limitation of this mouse system is the strong selective pressure to maintain IN-TP function, either directly by recombination or indirectly using mechanisms such as transcomplementation with expressed murine endogenous elements. Alternative animal models or assay systems (189, 190) would be beneficial to assess the full potential of MLV IN TP$^-$ as a vector in the absence of endogenous elements influencing integration preferences. The results also indicate that removing TP was not sufficient to redirect all integrations away from active promoters and strong enhancers, or to eliminate the stochastic events that can select for oncogenic activation. Ultimately, a modified vector that combines many different strategies could decrease genotoxicity and overcome current limitations for clinical applications. Self-inactivating (SIN) vectors eliminate strong viral enhancers (3, 191). For keratinocytes, MLV SIN vectors target less cell-growth related genes, TSS and epigenetically defined promoters. (191). Peptide motifs and protein domains that interact with heterochomatin regions can be inserted into retroviral genome and redirect integrations to less transcriptionally active genomic regions (125, 126).

Currently, our lab has developed different approaches to retarget integration. One strategy involves insertions of alternative chomatin-binding peptides in place of IN TP (Appendix 5). Another approach is replacing MLV IN CCD residues with PFV IN residues interacting with nucleosomes based on structure-based alignment. Structure of nucleosome bound PFV intasome shows interaction between residues located in loop regions between β1- β2 sheets and α4-α5 helices of the PFV IN CCD and H2A C-terminal helix and H2B (Appendix 6). This could possibly be a PFV specific interaction with H2A-H2B and H3 histones that could influence target-site recognition. These modifications are in MLV expressing constructs and expressed viral titer similar to MLV IN XN. Preliminary data on integration profile in 293mCAT cell for some MLV IN CCD (GLY→PSQ) and MLV IN CCD (GSRD→RPTD) have been collected. Along with these modifications, three copies of the p12 chromatin tethering domain of MLV has been added at the C-terminal end of IN for integration retargeting. The expected integrations for these modified MLV IN constructs are to regions that have low transcription activity thus reducing any potential oncogenic activation. Combination of these approaches can potentially decrease further genotoxicity when used for future clinical applications.

Bibliography

1. Vogt PK. Historical introduction to the general properties of retroviruses. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997.
2. Maetzig T, Baum C, Schambach A. Retroviral protein transfer: Falling apart to make an impact. Current Gene Therapy. 2012;12(5):389-409.
3. Maetzig T, Galla M, Baum C, Schambach A. Gammaretroviral vectors: Biology, technology and application. Viruses. 2011;3(6):677-713.
4. Hallberg B, Grundstrom T. Tissue specific sequence motifs in the enhancer of the leukaemogenic mouse retrovirus sl3-3. Nucleic Acids Res. 1988;16(13):5927-44.
5. Boral AL, Okenquist SA, Lenz J. Identification of the SL3-3 virus enhancer core as a T-lymphoma cell-specific element. J Virol. 1989;63(1):76-84.
6. Vile RG, Diaz RM, Miller N, Mitchell S, Tuszyanski A, Russell SJ. Tissue-specific gene expression from Mo-MLV retroviral vectors with hybrid LTRs containing the murine tyrosinase enhancer/promoter. Virology. 1995;214(1):307-13.
7. Fan H. Leukemogenesis by Moloney murine leukemia virus: A multistep process. Trends Microbiol. 1997;5(2):74-82.
8. Vile RG, Sunassee K, Diaz RM. Strategies for achieving multiple layers of selectivity in gene therapy. Mol Med Today. 1998;4(2):84-92.
9. Chow SA, Vincent KA, Ellison V, Brown PO. Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. Science. 1992;255(5045):723-6.
10. Zhou M, Deng L, Kashanchi F, Brady JN, Shatkin AJ, Kumar A. The tat/tar-dependent phosphorylation of RNA polymerase II C-terminal domain stimulates cotranscriptional capping of HIV-1 mRNA. Proc Natl Acad Sci U S A. 2003;100(22):12666-71.
11. Levin JG, Seidman JG. Selective packaging of host tRNA's by murine leukemia virus particles does not require genomic RNA. J Virol. 1979;29(1):328-35.
12. Peters G, Harada F, Dahlberg JE, Panet A, Haseltine WA, Baltimore D. Low-molecular-weight RNAs of Moloney murine leukemia virus: Identification of the primer for RNA-directed DNA synthesis. J Virol. 1977;21(3):1031-41.
13. Berkowitz RD, Ohagen A, Hoglund S, Goff SP. Retroviral nucleocapsid domains mediate the specific recognition of genomic viral RNAs by chimeric gag polyproteins during RNA packaging in vivo. J Virol. 1995;69(10):6445-56.
14. Bannert N, Fiebig U, Hohn O. Retroviral particles, proteins and genomes. In: Kurth R, Bannert N, editors. Retroviruses: Molecular biology, genomics and pathogenesis. 10. Robert Koch-Institut, 13353 Berlin, Germany: Caister Academic Press; 2010. p. 454.
15. Feher A, Boross P, Sperka T, Miklossy G, Kadas J, Bagossi P, Oroszlan S, Weber IT, Tozser J. Characterization of the murine leukemia virus protease and its comparison with the human immunodeficiency virus type 1 protease. J Gen Virol. 2006;87(Pt 5):1321-30.
16. Yoshinaka Y, Katoh I, Copeland TD, Oroszlan S. Murine leukemia-virus protease is encoded by the *gag-pol* gene and is synthesized through suppression of an amber termination codon. Proc Natl Acad Sci U S A. 1985;82(6):1618-22.
17. Orlova M, Yueh A, Leung J, Goff SP. Reverse transcriptase of Moloney murine leukemia virus binds to eukaryotic release factor 1 to modulate suppression of translational termination. Cell. 2003;115(3):319-31.

18. Jacks T, Varmus HE. Expression of the rous sarcoma virus *pol* gene by ribosomal frameshifting. Science. 1985;230(4731):1237-42.
19. Jacks T, Townsley K, Varmus HE, Majors J. Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus *gag*-related polyproteins. Proc Natl Acad Sci U S A. 1987;84(12):4298-302.
20. Giedroc DP, Cornish PV. Frameshifting RNA pseudoknots: Structure and mechanism. Virus Res. 2009;139(2):193-208.
21. Prats A-C, De Billy G, Wang P, Darlix J-L. CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus. J Mol Biol. 1989;205(2):363-72.
22. Pillemer EA, Kooistra DA, Witte ON, Weissman IL. Monoclonal antibody to the amino-terminal I sequence of murine leukemia virus glycosylated Gag polyproteins demonstrates their unusual orientation in the cell membrane. J Virol. 1986;57(2):413.
23. Nitta T, Tam R, Kim JW, Fan H. The cellular protein La functions in enhancement of virus release through lipid rafts facilitated by murine leukemia virus glycosylated Gag. MBio. 2011;2(1):e00341-10.
24. Schwartzberg P, Colicelli J, Goff SP. Deletion mutants of Moloney murine leukemia virus which lack glycosylated Gag protein are replication competent. J Virol. 1983;46(2):538-46.
25. Low A, Datta S, Kuznetsov Y, Jahid S, Kothari N, McPherson A, Fan H. Mutation in the glycosylated gag protein of murine leukemia virus results in reduced in vivo infectivity and a novel defect in viral budding or release. J Virol. 2007;81(8):3685-92.
26. Thomas JA, Gorelick RJ. Nucleocapsid protein function in early infection processes. Virus Res. 2008;134(1-2):39-63.
27. Chamontin C, Yu B, Racine PJ, Darlix JL, Mougel M. MoMLV and HIV-1 nucleocapsid proteins have a common role in genomic RNA packaging but different in late reverse transcription. PLoS One. 2012;7(12):e51534.
28. Demirov DG, Freed EO. Retrovirus budding. Virus Res. 2004;106(2):87-102.
29. Elis E, Ehrlich M, Prizan-Ravid A, Laham-Karam N, Bacharach E. p12 tethers the murine leukemia virus pre-integration complex to mitotic chromosomes. PLoS Pathog. 2012;8(12):e1003103.
30. Schneider WM, Brzezinski JD, Aiyer S, Malani N, Gyuricza M, Bushman FD, Roth MJ. Viral DNA tethering domains complement replication-defective mutations in the p12 protein of MuLV Gag. Proc Natl Acad Sci U S A. 2013;110(23):9487-92.
31. Wight DJ, Boucherit VC, Nader M, Allen DJ, Taylor IA, Bishop KN. The gammaretroviral p12 protein has multiple domains that function during the early stages of replication. Retrovirology. 2012;9(1):83.
32. Brzezinski JD, Felkner R, Modi A, Liu M, Roth MJ. Phosphorylation requirement of murine leukemia virus p12. J Virol. 2016;90(24):11208-19.
33. Wanaguru M, Barry DJ, Benton DJ, O'Reilly NJ, Bishop KN. Murine leukemia virus p12 tethers the capsid-containing pre-integration complex to chromatin by binding directly to host nucleosomes in mitosis. PLoS Pathog. 2018;14(6):e1007117.
34. Brzezinski JD, Modi A, Liu M, Roth MJ. Repression of the chromatin-tethering domain of murine leukemia virus p12. J Virol. 2016;90(24):11197-207.
35. Li M, Rasulova F, Melnikov EE, Rotanova TV, Gustchina A, Maurizi MR, Wlodawer A. Crystal structure of the N-terminal domain of *E. coli* Lon protease. Protein Sci. 2005;14(11):2895-900.

36. Tanese N, Goff SP. Domain structure of the Moloney murine leukemia virus reverse transcriptase: Mutational analysis and separate expression of the DNA polymerase and RNase H activities. Proc Natl Acad Sci U S A. 1988;85(6):1777-81.

37. Roth MJ, Tanese N, Goff SP. Purification and characterization of murine retroviral reverse transcriptase expressed in *Escherichia coli*. J Biol Chem. 1985;260(16):9326-35.

38. Moelling K. Characterization of reverse transcriptase and RNase H from Friend-murine leukemia virus. Virology. 1974;62(1):46-59.

39. Lesbats P, Engelman AN, Cherepanov P. Retroviral DNA integration. Chem Rev. 2016;116(20):12730-57.

40. Grandgenett DP, Pandey KK, Bera S, Aihara H. Multifunctional facets of retrovirus integrase. World J Biol Chem. 2015;6(3):83-94.

41. Ng VL, Wood TG, Arlinghaus RB. Processing of the gene products of Moloney murine leukaemia virus. J Gen Virol. 1982;59(Pt 2):329-43.

42. Lavillette D, Maurice M, Roche C, Russell SJ, Sitbon M, Cosset FL. A proline-rich motif downstream of the receptor binding domain modulates conformation and fusogenicity of murine retroviral envelopes. J Virol. 1998;72(12):9955-65.

43. Bae Y, Kingsman SM, Kingsman AJ. Functional dissection of the Moloney murine leukemia virus envelope protein gp70. J Virol. 1997;71(3):2092-9.

44. Rein A, Mirro J, Haynes JG, Ernst SM, Nagashima K. Function of the cytoplasmic domain of a retroviral transmembrane protein: p15e-p2e cleavage activates the membrane fusion capability of the murine leukemia virus env protein. J Virol. 1994;68(3):1773-81.

45. Jones JS, Risser R. Cell fusion induced by the murine leukemia virus envelope glycoprotein. J Virol. 1993;67(1):67-74.

46. Wallin M, Ekstrom M, Garoff H. Isomerization of the intersubunit disulphide-bond in Env controls retrovirus fusion. EMBO J. 2004;23(1):54-65.

47. Roe T, Reynolds TC, Yu G, Brown PO. Integration of murine leukemia virus DNA depends on mitosis. EMBO J. 1993;12(5):2099-108.

48. Lin CW, Engelman A. The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. Journal of Virology. 2003;77(8):5030-6.

49. Kraunus J, Zychlinski D, Heise T, Galla M, Bohne J, Baum C. Murine leukemia virus regulates alternative splicing through sequences upstream of the 5' splice site. J Biol Chem. 2006;281(49):37381-90.

50. Andrawiss M, Takeuchi Y, Hewlett L, Collins M. Murine leukemia virus particle assembly quantitated by fluorescence microscopy: Role of Gag-Gag interactions and membrane association. J Virol. 2003;77(21):11651-60.

51. Rein A, McClure MR, Rice NR, Luftig RB, Schultz AM. Myristylation site in pr65Gag is essential for virus particle formation by Moloney murine leukemia virus. Proc Natl Acad Sci U S A. 1986;83(19):7246-50.

52. Schultz AM, Rein A. Unmyristylated Moloney murine leukemia virus pr65Gag is excluded from virus assembly and maturation events. J Virol. 1989;63(5):2370-3.

53. Hadravova R, de Marco A, Ulbrich P, Stokrova J, Dolezal M, Pichova I, Ruml T, Briggs JA, Rumlova M. In vitro assembly of virus-like particles of a gammaretrovirus, the murine leukemia virus XMRV. J Virol. 2012;86(3):1297-306.

54. Yeager M, Wilson-Kubalek EM, Weiner SG, Brown PO, Rein A. Supramolecular organization of immature and mature murine leukemia virus revealed by electron

cryo-microscopy: Implications for retroviral assembly mechanisms. Proc Natl Acad Sci U S A. 1998;95(13):7299-304.

55. Ao Z, Fowke KR, Cohen EA, Yao X. Contribution of the C-terminal tri-lysine regions of human immunodeficiency virus type 1 integrase for efficient reverse transcription and viral DNA nuclear import. Retrovirology. 2005;2:62.

56. Ikeda T, Nishitsuji H, Zhou X, Nara N, Ohashi T, Kannagi M, Masuda T. Evaluation of the functional involvement of human immunodeficiency virus type 1 integrase in nuclear import of viral cDNA during acute infection. J Virol. 2004;78(21):11563-73.

57. Lu R, Limon A, Ghory HZ, Engelman A. Genetic analyses of DNA-binding mutants in the catalytic core domain of human immunodeficiency virus type 1 integrase. J Virol. 2005;79(4):2493-505.

58. Engelman A, Englund G, Orenstein JM, Martin MA, Craigie R. Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. J Virol. 1995;69(5):2729-36.

59. Aiyer S, Rossi P, Malani N, Schneider WM, Chandar A, Bushman FD, Montelione GT, Roth MJ. Structural and sequencing analysis of local target DNA recognition by MLV integrase. Nucleic Acids Res. 2015;43(11):5647-63.

60. Aiyer S, Swapna GV, Malani N, Aramini JM, Schneider WM, Plumb MR, Ghanem M, Larue RC, Sharma A, Studamire B, Kvaratskhelia M, Bushman FD, Montelione GT, Roth MJ. Altering murine leukemia virus integration through disruption of the integrase and BET protein family interaction. Nucleic Acids Res. 2014;42(9):5917-28.

61. Guan R, Aiyer S, Cote ML, Xiao R, Jiang M, Acton TB, Roth MJ, Montelione GT. X-ray crystal structure of the N-terminal region of Moloney murine leukemia virus integrase and its implications for viral DNA recognition. Proteins. 2017;85(4):647-56.

62. Yin Z, Shi K, Banerjee S, Pandey KK, Bera S, Grandgenett DP, Aihara H. Crystal structure of the rous sarcoma virus intasome. Nature. 2016;530(7590):362-6.

63. Ballandras-Colas A, Brown M, Cook NJ, Dewdney TG, Demeler B, Cherepanov P, Lyumkis D, Engelman AN. Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. Nature. 2016;530(7590):358-61.

64. Maskell DP, Renault L, Serrao E, Lesbats P, Matadeen R, Hare S, Lindemann D, Engelman AN, Costa A, Cherepanov P. Structural basis for retroviral integration into nucleosomes. Nature. 2015;523(7560):366-9.

65. Passos DO, Li M, Yang R, Rebensburg SV, Ghirlando R, Jeon Y, Shkriabai N, Kvaratskhelia M, Craigie R, Lyumkis D. Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. Science. 2017;355(6320):89-92.

66. Grawenhoff J, Engelman AN. Retroviral integrase protein and intasome nucleoprotein complex structures. World J Biol Chem. 2017;8(1):32-44.

67. Puglia J, Wang T, Smith-Snyder C, Cote M, Scher M, Pelletier JN, John S, Jonsson CB, Roth MJ. Revealing domain structure through linker-scanning analysis of the murine leukemia virus (MuLV) RNase H and MuLV and human immunodeficiency virus type 1 integrase proteins. J Virol. 2006;80(19):9497-510.

68. Poletti V, Mavilio F. Interactions between retroviruses and the host cell genome. Mol Ther Methods Clin Dev. 2018;8:31-41.

69. De Rijck J, de Kogel C, Demeulemeester J, Vets S, El Ashkar S, Malani N, Bushman FD, Landuyt B, Husson SJ, Busschots K, Gijsbers R, Debyser Z. The BET family of proteins targets Moloney murine leukemia virus integration near transcription start sites. Cell Rep. 2013;5(4):886-94.

70. El Ashkar S, De Rijck J, Demeulemeester J, Vets S, Madlala P, Cermakova K, Debyser Z, Gijsbers R. BET-independent MLV-based vectors target away from promoters and regulatory elements. Mol Ther Nucleic Acids. 2014;3:e179.

71. Villanueva RA, Jonsson CB, Jones J, Georgiadis MM, Roth MJ. Differential multimerization of Moloney murine leukemia virus integrase purified under nondenaturing conditions. Virology. 2003;316(1):146-60.

72. Yang F, Leon O, Greenfield NJ, Roth MJ. Functional interactions of the hhcc domain of Moloney murine leukemia virus integrase revealed by nonoverlapping complementation and zinc-dependent dimerization. J Virol. 1999;73(3):1809-17.

73. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. Cell. 2002;110(4):521-9.

74. Serrao E, Engelman AN. Sites of retroviral DNA integration: From basic research to clinical applications. Crit Rev Biochem Mol Biol. 2016;51(1):26-42.

75. LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res. 2014;42(7):4257-69.

76. De Ravin SS, Su L, Theobald N, Choi U, Macpherson JL, Poidinger M, Symonds G, Pond SM, Ferris AL, Hughes SH, Malech HL, Wu X. Enhancers are major targets for murine leukemia virus vector integration. J Virol. 2014;88(8):4504-13.

77. Engelman A, Cherepanov P. Retroviral integrase structure and DNA recombination mechanism. Microbiology Spectrum. 2014;2(6):1011-33.

78. Kvaratskhelia M, Sharma A, Larue RC, Serrao E, Engelman A. Molecular mechanisms of retroviral integration site selection. Nucleic Acids Res. 2014;42(16):10209-25.

79. Serrao E, Ballandras-Colas A, Cherepanov P, Maertens GN, Engelman AN. Key determinants of target DNA recognition by retroviral intasomes. Retrovirology. 2015;12:39.

80. Cherepanov P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. Nucleic Acids Res. 2007;35(1):113.

81. Cherepanov P, Ambrosio AL, Rahman S, Ellenberger T, Engelman A. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. Proc Natl Acad Sci U S A. 2005;102:17308.

82. Cherepanov P, Devroe E, Silver PA, Engelman A. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. J Biol Chem. 2004;279(47):48883.

83. Cherepanov P, Maertens G, Proost P, Devreese B, Van Beeumen J, Engelborghs Y, De Clercq E, Debyser Z. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. J Biol Chem. 2003;278(1):372-81.

84. Winans S, Larue RC, Abraham CM, Shkriabai N, Skopp A, Winkler D, Kvaratskhelia M, Beemon KL. The fact complex promotes avian leukosis virus DNA integration. J Virol. 2017;91(7):e00082-17.

85. Matysiak J, Lesbats P, Mauro E, Lapaillerie D, Dupuy JW, Lopez AP, Benleulmi MS, Calmels C, Andreola ML, Ruff M, Llano M, Delelis O, Lavigne M, Parissi V. Modulation of chromatin structure by the fact histone chaperone complex regulates HIV-1 integration. Retrovirology. 2017;14(1):39.

86. Stathis A, Bertoni F. BET proteins as targets for anticancer treatment. Cancer Discov. 2018;8(1):24-36.

87. Taniguchi Y. The bromodomain and extra-terminal domain (BET) family: Functional anatomy of BET paralogous proteins. Int J Mol Sci. 2016;17(11).

88. Berns A, Adams DJ, de Ridder J, de Jong J, Jonkers J, van der Weyden L, van Lohuizen M, van Uitert M, Sun N, Wessels LFA. Computational identification of insertional mutagenesis targets for cancer gene discovery. Nucleic Acids Research. 2011;39(15):e105-e.

89. Kool J, Uren AG, Martins CP, Sie D, de Ridder J, Turner G, van Uitert M, Matentzoglu K, Lagcher W, Krimpenfort P, Gadiot J, Pritchard C, Lenz J, Lund AH, Jonkers J, Rogers J, Adams DJ, Wessels L, Berns A, van Lohuizen M. Insertional mutagenesis in mice deficient for p15ink4b, p16ink4a, p21cip1, and p27kip1 reveals cancer gene interactions and correlations with tumor phenotypes. Cancer Res. 2010;70(2):520-31.

90. Huser CA, Gilroy KL, de Ridder J, Kilbey A, Borland G, Mackay N, Jenkins A, Bell M, Herzyk P, van der Weyden L, Adams DJ, Rust AG, Cameron E, Neil JC. Insertional mutagenesis and deep profiling reveals gene hierarchies and a MYC/p53-dependent bottleneck in lymphomagenesis. PLoS Genet. 2014;10(2):e1004167.

91. Steffen D. Proviruses are adjacent to c-myc in some murine leukemia virus-inducedlymphomas. Proc Natl Acad Sci U S A. 1984;81(7):2097-101.

92. Fan H, Johnson C. Insertional oncogenesis by non-acute retroviruses: Implications for gene therapy. Viruses. 2011;3(4):398-422.

93. Ranzani M, Annunziato S, Adams DJ, Montini E. Cancer gene discovery: Exploiting insertional mutagenesis. Mol Cancer Res. 2013;11(10):1141-58.

94. Cavazza A, Moiani A, Mavilio F. Mechanisms of retroviral integration and mutagenesis. Hum Gene Ther. 2013;24(2):119-31.

95. Touw IP, Erkeland SJ. Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia. Mol Ther. 2007;15(1):13-9.

96. Cavazzana-Calvo M, Payen E, Negre O, Wang G, Hehir K, Fusil F, Down J, Denaro M, Brady T, Westerman K, Cavallesco R, Gillet-Legrand B, Caccavelli L, Sgarra R, Maouche-Chretien L, Bernaudin F, Girot R, Dorazio R, Mulder GJ, Polack A, Bank A, Soulier J, Larghero J, Kabbara N, Dalle B, Gourmel B, Socie G, Chretien S, Cartier N, Aubourg P, Fischer A, Cornetta K, Galacteros F, Beuzard Y, Gluckman E, Bushman F, Hacein-Bey-Abina S, Leboulch P. Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. Nature. 2010;467(7313):318-22.

97. Maruggi G, Porcellini S, Facchini G, Perna SK, Cattoglio C, Sartori D, Ambrosi A, Schambach A, Baum C, Bonini C, Bovolenta C, Mavilio F, Recchia A. Transcriptional enhancers induce insertional gene deregulation independently from the vector type and design. Mol Ther. 2009;17(5):851-6.

98. Recchia A, Bonini C, Magnani Z, Urbinati F, Sartori D, Muraro S, Tagliafico E, Bondanza A, Stanghellini MT, Bernardi M, Pescarollo A, Ciceri F, Bordignon C, Mavilio F. Retroviral vector integration deregulates gene expression but has no consequence on the biology and function of transplanted T cells. Proc Natl Acad Sci U S A. 2006;103(5):1457-62.

99. Neil JC, Gilroy K, Borland G, Hay J, Terry A, Kilbey A. The Runx genes as conditional oncogenes: Insights from retroviral targeting and mouse models. In: Groner Y, Ito Y, Liu P, Neil JC, Speck NA, van Wijnen A, editors. Runx proteins in development and cancer. Singapore: Springer Singapore; 2017. p. 247-64.

100. Blyth K, Vaillant F, Hanlon L, Mackay N, Bell M, Jenkins A, Neil JC, Cameron ER. *Runx2* and *MYC* collaborate in lymphoma development by suppressing apoptotic and growth arrest pathways in vivo. Cancer Res. 2006;66(4):2195-201.

101. Kozak CA. Origins of the endogenous and infectious laboratory mouse gammaretroviruses. Viruses. 2014;7(1):1-26.

102. Greenwood AD, Ishida Y, O'Brien SP, Roca AL, Eiden MV. Transmission, evolution, and endogenization: Lessons learned from recent retroviral invasions. Microbiol Mol Biol Rev. 2018;82(1):e00044-17.

103. Stoye JP, Coffin JM. The four classes of endogenous murine leukemia virus: Structural relationships and potential for recombination. J Virol. 1987;61(9):2659-69.

104. Giovannini D, Touhami J, Charnet P, Sitbon M, Battini JL. Inorganic phosphate export by the retrovirus receptor Xpr1 in metazoans. Cell Rep. 2013;3(6):1866-73.

105. Anderson JA, Teufel RJ, 2nd, Yin PD, Hu WS. Correlated template-switching events during minus-strand DNA synthesis: A mechanism for high negative interference during retroviral recombination. J Virol. 1998;72(2):1186-94.

106. Hu WS, Temin HM. Retroviral recombination and reverse transcription. Science. 1990;250(4985):1227-33.

107. Stoye JP. Endogenous retroviruses: Still active after all these years? Current Biology. 2001;11(22):R914-R6.

108. Zhuang J, Mukherjee S, Ron Y, Dougherty JP. High rate of genetic recombination in murine leukemia virus: Implications for influencing proviral ploidy. J Virol. 2006;80(13):6706-11.

109. Onafuwa-Nuga A, Telesnitsky A. The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. Microbiol Mol Biol Rev. 2009;73(3):451-80, Table of Contents.

110. Bamunusinghe D, Liu Q, Plishka R, Dolan MA, Skorski M, Oler AJ, Yedavalli VRK, Buckler-White A, Hartley JW, Kozak CA. Recombinant origins of pathogenic and nonpathogenic mouse gammaretroviruses with polytropic host range. J Virol. 2017;91(21):e00855-17.

111. Stoye JP, Moroni C, Coffin JM. Virological events leading to spontaneous AKR thymomas. J Virol. 1991;65(3):1273-85.

112. Ott D, Friedrich R, Rein A. Sequence analysis of amphotropic and loal murine leukemia viruses: Close relationship to mink cell focus-inducing viruses. J Virol. 1990;64(2):757-66.

113. Kozak C, Rowe WP. Genetic mapping of xenotropic leukemia virus-inducing loci in two mouse strains. Science. 1978;199(4336):1448-9.

114. Kohn DB. Historical perspective on the current renaissance for hematopoietic stem cell gene therapy. Hematol Oncol Clin North Am. 2017;31(5):721-35.

115. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, Lim A, Osborne CS, Pawliuk R, Morillon E, Sorensen R, Forster A, Fraser P, Cohen JI, de Saint Basile G, Alexander I, Wintergerst U, Frebourg T, Aurias A, Stoppa-Lyonnet D, Romana S, Radford-Weiss I, Gross F, Valensi F, Delabesse E, Macintyre E, Sigaux F, Soulier J, Leiva LE, Wissler M, Prinz C, Rabbitts TH, Le Deist F, Fischer A, Cavazzana-Calvo M. Lmo2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science. 2003;302(5644):415-9.

116. Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier J, Lim A, Morillon E, Clappier E, Caccavelli L, Delabesse E, Beldjord K, Asnafi V, MacIntyre E, Dal Cortivo L, Radford I, Brousse N, Sigaux F, Moshous D, Hauer J, Borkhardt A,

Belohradsky BH, Wintergerst U, Velez MC, Leiva L, Sorensen R, Wulffraat N, Blanche S, Bushman FD, Fischer A, Cavazzana-Calvo M. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J Clin Invest. 2008;118(9):3132-42.

117. Hacein-Bey-Abina S, Hauer J, Lim A, Picard C, Wang GP, Berry CC, Martinache C, Rieux-Laucat F, Latour S, Belohradsky BH, Leiva L, Sorensen R, Debre M, Casanova JL, Blanche S, Durandy A, Bushman FD, Fischer A, Cavazzana-Calvo M. Efficacy of gene therapy for X-linked severe combined immunodeficiency. N Engl J Med. 2010;363(4):355-64.

118. Stein S, Ott MG, Schultze-Strasser S, Jauch A, Burwinkel B, Kinner A, Schmidt M, Kramer A, Schwable J, Glimm H, Koehl U, Preiss C, Ball C, Martin H, Gohring G, Schwarzwaelder K, Hofmann WK, Karakaya K, Tchatchou S, Yang R, Reinecke P, Kuhlcke K, Schlegelberger B, Thrasher AJ, Hoelzer D, Seger R, von Kalle C, Grez M. Genomic instability and myelodysplasia with monosomy 7 consequent to Evi1 activation after gene therapy for chronic granulomatous disease. Nat Med. 2010;16(2):198-204.

119. Braun CJ, Boztug K, Paruzynski A, Witzel M, Schwarzer A, Rothe M, Modlich U, Beier R, Gohring G, Steinemann D, Fronza R, Ball CR, Haemmerle R, Naundorf S, Kuhlcke K, Rose M, Fraser C, Mathias L, Ferrari R, Abboud MR, Al-Herz W, Kondratenko I, Marodi L, Glimm H, Schlegelberger B, Schambach A, Albert MH, Schmidt M, von Kalle C, Klein C. Gene therapy for wiskott-aldrich syndrome--long-term efficacy and genotoxicity. Sci Transl Med. 2014;6(227):227ra33.

120. Wu C, Dunbar CE. Stem cell gene therapy: The risks of insertional mutagenesis and approaches to minimize genotoxicity. Front Med. 2011;5(4):356-71.

121. McCormack MP, Rabbitts TH. Activation of the t-cell oncogene Lmo2 after gene therapy for X-linked severe combined immunodeficiency. New England Journal of Medicine. 2004;350(9):913-22.

122. Wang GP, Berry CC, Malani N, Leboulch P, Fischer A, Hacein-Bey-Abina S, Cavazzana-Calvo M, Bushman FD. Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. Blood. 2010;115(22):4356-66.

123. Aiuti A, Cassani B, Andolfi G, Mirolo M, Biasco L, Recchia A, Urbinati F, Valacca C, Scaramuzza S, Aker M, Slavin S, Cazzola M, Sartori D, Ambrosi A, Di Serio C, Roncarolo MG, Mavilio F, Bordignon C. Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. J Clin Invest. 2007;117(8):2233-40.

124. Hacein-Bey-Abina S, Pai SY, Gaspar HB, Armant M, Berry CC, Blanche S, Bleesing J, Blondeau J, de Boer H, Buckland KF, Caccavelli L, Cros G, De Oliveira S, Fernandez KS, Guo D, Harris CE, Hopkins G, Lehmann LE, Lim A, London WB, van der Loo JC, Malani N, Male F, Malik P, Marinovic MA, McNicol AM, Moshous D, Neven B, Oleastro M, Picard C, Ritz J, Rivat C, Schambach A, Shaw KL, Sherman EA, Silberstein LE, Six E, Touzot F, Tsytsykova A, Xu-Bayford J, Baum C, Bushman FD, Fischer A, Kohn DB, Filipovich AH, Notarangelo LD, Cavazzana M, Williams DA, Thrasher AJ. A modified gamma-retrovirus vector for X-linked severe combined immunodeficiency. N Engl J Med. 2014;371(15):1407-17.

125. El Ashkar S, Van Looveren D, Schenk F, Vranckx LS, Demeulemeester J, De Rijck J, Debyser Z, Modlich U, Gijsbers R. Engineering next-generation BET-independent MLV vectors for safer gene therapy. Mol Ther Nucleic Acids. 2017;7:231-45.

126. Nam JS, Lee JE, Lee KH, Yang Y, Kim SH, Bae GU, Noh H, Lim KI. Shifting retroviral vector integrations away from transcriptional start sites via DNA-binding protein domain insertion into integrase. Mol Ther Methods Clin Dev. 2019;12:58-70.
127. O'Reilly L, Roth MJ. Second-site changes affect viability of amphotropic/ecotropic chimeric enveloped murine leukemia viruses. J Virol. 2000;74(2):899-913.
128. Felkner RH, Roth MJ. Mutational analysis of the N-linked glycosylation sites of the SU envelope protein of Moloney murine leukemia virus. J Virol. 1992;66(7):4258-64.
129. Roth MJ. Mutational analysis of the carboxyl terminus of the Moloney murine leukemia virus integration protein. J Virol. 1991;65(4):2141-5.
130. Schneider WM, Wu DT, Amin V, Aiyer S, Roth MJ. Mulv IN mutants responsive to HDAC inhibitors enhance transcription from unintegrated retroviral DNA. Virology. 2012;426(2):188-96.
131. McCutchan JH, Pagano JS. Enhancement of the infectivity of simian virus 40 deoxyribonucleic acid with diethylaminoethyl-dextran. J Natl Cancer Inst. 1968;41:351-7.
132. Wu DT, Aiyer S, Villanueva RA, Roth MJ. Development of an enzyme-linked immunosorbent assay based on the murine leukemia virus p30 capsid protein. J Virol Methods. 2013;193(2):332-6.
133. Ting YT, Wilson CA, Farrell KB, Chaudry GJ, Eiden MV. Simian sarcoma-associated virus fails to infect chinese hamster cells despite the presence of functional gibbon ape leukemia virus receptors. J Virol. 1998;72(12):9453-8.
134. Valdivieso-Torres L, Sarangi A, Whidby J, Marcotrigiano J, Roth MJ. Role of cysteines in stabilizing the randomized receptor binding domains within feline leukemia virus envelope proteins. J Virol. 2015;90(6):2971-80.
135. Peredo C, O'Reilly L, Gray K, Roth MJ. Characterization of chimeras between the ecotropic Moloney murine leukemia virus and the amphotropic 4070a envelope proteins. J Virol. 1996;70(5):3142-52.
136. Tanese N, Roth MJ, Goff SP. Analysis of retroviral *pol* gene products with antisera raised against fusion proteins produced in *Escherichia coli*. J Virol. 1986;59(2):328-40.
137. Acton TB, Xiao R, Anderson S, Aramini J, Buchwald WA, Ciccosanti C, Conover K, Everett J, Hamilton K, Huang YJ, Janjua H, Kornhaber G, Lau J, Lee DY, Liu G, Maglaqui M, Ma L, Mao L, Patel D, Rossi P, Sahdev S, Shastry R, Swapna GV, Tang Y, Tong S, Wang D, Wang H, Zhao L, Montelione GT. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. Methods Enzymol. 2011;493:21-60.
138. Khan AG, Whidby J, Miller MT, Scarborough H, Zatorski AV, Cygan A, Price AA, Yost SA, Bohannon CD, Jacob J, Grakoui A, Marcotrigiano J. Structure of the core ectodomain of the Hepatitis C virus envelope glycoprotein 2. Nature. 2014;509(7500):381-4.
139. Zheng J, Wang C, Chang MR, Devarkar SC, Schweibenz B, Crynen GC, Garcia-Ordonez RD, Pascal BD, Novick SJ, Patel SS, Marcotrigiano J, Griffin PR. Hdx-ms reveals dysregulated checkpoints that compromise discrimination against self RNA during RIG-I mediated autoimmunity. Nat Commun. 2018;9(1):5366.
140. Sharma S, Zheng H, Huang YJ, Ertekin A, Hamuro Y, Rossi P, Tejero R, Acton TB, Xiao R, Jiang M, Zhao L, Ma LC, Swapna GV, Aramini JM, Montelione GT. Construct optimization for protein NMR structure analysis using amide hydrogen/deuterium exchange mass spectrometry. Proteins. 2009;76(4):882-94.

141. Stewart M, Mackay N, Hanlon L, Blyth K, Scobie L, Cameron E, Neil JC. Insertional mutagenesis reveals progression genes and checkpoints in *MYC/Runx2* lymphomas. Cancer Res. 2007;67(11):5126-33.

142. Uphoff CC, Lange S, Denkmann SA, Garritsen HS, Drexler HG. Prevalence and characterization of murine leukemia virus contamination in human cell lines. PLoS One. 2015;10(4):e0125622.

143. Zheng H, Jia H, Shankar A, Heneine W, Switzer WM. Detection of murine leukemia virus or mouse DNA in commercial RT-PCR reagents and human dnas. PLoS One. 2011;6(12):e29050.

144. Serrao E, Cherepanov P, Engelman AN. Amplification, next-generation sequencing, and genomic DNA mapping of retroviral integration sites. J Vis Exp. 2016(109).

145. Wang GG, Calvo KR, Pasillas MP, Sykes DB, Hacker H, Kamps MP. Quantitative production of macrophages or neutrophils ex vivo using conditional Hoxb8. Nat Methods. 2006;3(4):287-93.

146. Achuthan V, Perreira JM, Sowd GA, Puray-Chavez M, McDougall WM, Paulucci-Holthauzen A, Wu X, Fadel HJ, Poeschla EM, Multani AS, Hughes SH, Sarafianos SG, Brass AL, Engelman AN. Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. Cell Host Microbe. 2018;24(3):392-404 e8.

147. Kim D, Langmead B, Salzberg SL. Hisat: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357-60.

148. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078-9.

149. Quinlan AR, Hall IM. Bedtools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

150. Wickham H. Ggplot2. New York: Springer-Verlag; 2009.

151. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43-9.

152. Stoye JP, Coffin SM. Polymorphism of murine endogenous proviruses revealed by using virus class-specific oligonucleotide probes. J Virol. 1988;62:168-75.

153. Jia X, Lin X, Chen J. Linear and exponential tail-PCR: A method for efficient and quick amplification of flanking sequences adjacent to Tn5 transposon insertion sites. AMB Express. 2017;7(1):195.

154. Gupta SS, Maetzig T, Maertens GN, Sharif A, Rothe M, Weidner-Glunde M, Galla M, Schambach A, Cherepanov P, Schulz TF. Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. J Virol. 2013;87(23):12721-36.

155. Pei J, Kim BH, Grishin NV. Promals3d: A tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36(7):2295-300.

156. Robert X, Gouet P. Deciphering key features in protein structures with the new endscript server. Nucleic Acids Res. 2014;42(Web Server issue):W320-4.

157. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

158. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. Science. 2003;300(5626):1749-51.

159. Walters BT, Ricciuti A, Mayne L, Englander SW. Minimizing back exchange in the hydrogen exchange-mass spectrometry experiment. J Am Soc Mass Spectrom. 2012;23(12):2132-9.

160. Konermann L, Pan J, Liu YH. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. Chem Soc Rev. 2011;40(3):1224-34.

161. Ey P, Freeman N, Bela B, Li P, McInnes J. Nucleotide sequence of the murine leukemia virus amphotropic strain 4070A integrase (IN) coding region and comparative structural analysis of the inferred polypeptide. Arch Virol. 1997;142:1757-70.

162. Puglia J, Wang T, Smith-Snyder C, Cote M, Scher M, Pelletier J, John S, Jonsson C, Roth M. Revealing domain structure through linker-scanning analysis of the murine leukemia virus (MuLV) RNAse h and MuLV and human immunodeficiency virus type 1 integrase proteins. J Virol. 2006;80(19):9497-510.

163. Moiani A, Miccio A, Rizzi E, Severgnini M, Pellin D, Suerth JD, Baum C, De Bellis G, Mavilio F. Deletion of the LTR enhancer/promoter has no impact on the integration profile of MLV vectors in human hematopoietic progenitors. PLoS One. 2013;8(1):e55721.

164. Santoni FA, Hartley O, Luban J. Deciphering the code for retroviral integration target site selection. PLoS Comput Biol. 2010;6(11):e1001008.

165. LeRoy G, Chepelev I, DiMaggio PA, Blanco MA, Zee BM, Zhao K, Garcia BA. Proteogenomic characterization and mapping of nucleosomes decoded by brd and HP1 proteins. Genome Biol. 2012;13(8):R68.

166. Chapuy B, McKeown MR, Lin CY, Monti S, Roemer MG, Qi J, Rahl PB, Sun HH, Yeda KT, Doench JG, Reichert E, Kung AL, Rodig SJ, Young RA, Shipp MA, Bradner JE. Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. Cancer Cell. 2013;24(6):777-90.

167. Gilroy KL, Terry A, Naseer A, de Ridder J, Allahyar A, Wang W, Carpenter E, Mason A, Wong GK, Cameron ER, Kilbey A, Neil JC. Gamma-retrovirus integration marks cell type-specific cancer genes: A novel profiling tool in cancer genomics. PLoS One. 2016;11(4):e0154070.

168. Uren AG, Kool J, Berns A, van Lohuizen M. Retroviral insertional mutagenesis: Past, present and future. Oncogene. 2005;24(52):7656-72.

169. Baliji S, Liu Q, Kozak CA. Common inbred strains of the laboratory mouse that are susceptible to infection by mouse xenotropic gammaretroviruses and the human-derived retrovirus XMRV. J Virol. 2010;84(24):12841-9.

170. Lee KH, Horiuchi M, Itoh T, Greenhalgh DG, Cho K. Cerebellum-specific and age-dependent expression of an endogenous retrovirus with intact coding potential. Retrovirology. 2011;8:82.

171. McCubrey J, Risser R. Genetic interactions in induction of endogenous murine leukemia virus from low leukemic mice. Cell. 1982;28(4):881-8.

172. Williams AB, Schumacher B. p53 in the DNA-damage-repair process. Cold Spring Harb Perspect Med. 2016;6(5):a026070.

173. Bennett GR, Peters R, Wang XH, Hanne J, Sobol RW, Bundschuh R, Fishel R, Yoder KE. Repair of oxidative DNA base damage in the host genome influences the HIV integration site sequence preference. PLoS One. 2014;9(7):e103164.

174. Pattison JM, Wright JB, Cole MD. Retroviruses hijack chromatin loops to drive oncogene expression and highlight the chromatin architecture around proto-oncogenic loci. PLoS One. 2015;10(3):e0120256.

175. Nakagawa M, Tsuzuki S, Honma K, Taguchi O, Seto M. Synergistic effect of *Bcl2, Myc* and *Ccnd1* transforms mouse primary B cells into malignant cells. Haematologica. 2011;96(9):1318-26.

176. Junk DJ, Cipriano R, Stampfer M, Jackson MW. Constitutive Ccnd1/Cdk2 activity substitutes for p53 loss, or MYC or oncogenic Ras expression in the transformation of human mammary epithelial cells. PLoS One. 2013;8(2):e53776.

177. Arita K, Maeda-Kasugai Y, Ohshima K, Tsuzuki S, Suguro-Katayama M, Karube K, Yoshida N, Sugiyama T, Seto M. Generation of mouse models of lymphoid neoplasm using retroviral gene transduction of in vitro-induced germinal center B and T cells. Exp Hematol. 2013;41(8):731-41 e9.

178. Satou Y, Miyazato P, Ishihara K, Yaguchi H, Melamed A, Miura M, Fukuda A, Nosaka K, Watanabe T, Rowan AG, Nakao M, Bangham CR. The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. Proc Natl Acad Sci U S A. 2016;113(11):3054-9.

179. Melamed A, Yaguchi H, Miura M, Witkover A, Fitzgerald TW, Birney E, Bangham CR. The human leukemia virus HTLV-1 alters the structure and transcription of host chromatin in CIS. Elife. 2018;7.

180. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei CL, Liu ET, Wilczynski GM, Plewczynski D, Li G, Ruan Y. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015;163(7):1611-27.

181. Ong CT, Corces VG. Ctcf: An architectural protein bridging genome topology and function. Nat Rev Genet. 2014;15(4):234-46.

182. Ong CT, Corces VG. Enhancer function: New insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12(4):283-93.

183. Ozaki T, Wu D, Sugimoto H, Nagase H, Nakagawara A. Runt-related transcription factor 2 (Runx2) inhibits p53-dependent apoptosis through the collaboration with Hdac6 in response to DNA damage. Cell Death Dis. 2013;4(4):e610.

184. Westendorf JJ, Zaidi SK, Cascino JE, Kahler R, van Wijnen AJ, Lian JB, Yoshida M, Stein GS, Li X. Runx2 (Cbfa1, Aml-3) interacts with histone deacetylase 6 and represses the p21(Cip1/Waf1) promoter. Mol Cell Biol. 2002;22(22):7982-92.

185. Valenzuela-Fernandez A, Cabrero JR, Serrador JM, Sanchez-Madrid F. Hdac6: A key regulator of cytoskeleton, cell migration and cell-cell interactions. Trends Cell Biol. 2008;18(6):291-7.

186. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS. Integrative annotation of chromatin elements from encode data. Nucleic Acids Res. 2013;41(2):827-41.

187. McGinty RK, Tan S. Recognition of the nucleosome by chromatin factors and enzymes. Curr Opin Struct Biol. 2016;37:54-61.

188. Shaytan AK, Landsman D, Panchenko AR. Nucleosome adaptability conferred by sequence and structural variations in histone H2A-H2B dimers. Curr Opin Struct Biol. 2015;32:48-57.

189. Biasco L, Rothe M, Buning H, Schambach A. Analyzing the genotoxicity of retroviral vectors in hematopoietic cell gene therapy. Mol Ther Methods Clin Dev. 2018;8:21-30.

190. Rothe M, Schambach A, Biasco L. Safety of gene therapy: New insights to a puzzling case. Curr Gene Ther. 2014;14(6):429-36.

191. Cavazza A, Cocchiarella F, Bartholomae C, Schmidt M, Pincelli C, Larcher F, Mavilio F. Self-inactivating MLV vectors have a reduced genotoxic profile in human epidermal keratinocytes. Gene Ther. 2013;20(9):949-57.

Appendices

Appendix 1. *MYC/Runx2* mice infected with WT MLV and MLV IN XN viruses

	ID	Sex	DoB	DoD	Life	reason for death
WT MLV	GimWT inf-1	F	7/23/14	8/19/14	27	
	GimWT inf-2	F	7/23/14	8/26/14	34	
	GimWT inf-3	M	7/23/14	8/26/14	34	
	GimWT inf-4	M	7/23/14	8/26/14	34	
	GimWT inf-5	M	7/23/14	8/26/14	34	
	GimWT inf-6	F	7/29/14	8/28/14	30	
	GimWT inf-7	F	7/29/14	8/28/14	30	
	GimWT inf-13	F	7/29/14	9/1/14	34	
	GimWT inf-8	M	7/29/14	8/28/14	30	
	GimWT inf-9	M	7/29/14	8/29/14	31	
	GimWT inf-10	M	7/29/14	8/29/14	31	
	GimWT inf-11	M	7/29/14	9/1/14	34	
	GimWT inf-12	M	7/29/14	9/1/14	34	
	GimWT inf-14	M	7/29/14	9/2/14	35	
	GimWT inf-15	M	7/29/14	9/2/14	35	
	GimWT inf-16	M	7/29/14	9/2/14	35	
	GimWT inf-18	F	8/13/14	9/22/14	40	
	GimWT inf-17	M	8/13/14	9/18/14	36	
	GimWT inf-19	M	8/13/14	9/23/14	41	
	GimWT inf-20	M	8/13/14	9/30/14	48	
	GimWT inf-23	F	8/18/14	10/10/14	53	
	GimWT inf-24	F	8/18/14	10/16/14	59	
	GimWT inf-21	M	8/18/14	10/2/14	45	
	GimWT inf-22	M	8/18/14	10/8/14	51	
	GimWT inf-25	M	8/18/14	10/20/14	63	
	GimWT inf-26	M	8/18/14	10/20/14	63	
	GimWT inf-28	F	9/2/14	10/29/14	57	
	GimWT inf-29	F	9/2/14	10/29/14	57	
	GimWT inf-30	F	9/2/14	10/31/14	59	
	GimWT inf-27	M	9/2/14	10/24/14	52	
	GimWT inf-		9/2/14	9/11/14	9	thymus samples
	GimWT inf-		9/2/14	9/11/14	9	thymus samples
	GimWT inf-		9/2/14	9/11/14	9	thymus samples
MLV IN-XN	GimXN inf-1	F	7/24/14	9/2/14	40	
	GimXN inf-2	F	7/24/14	9/3/14	41	
	GimXN inf-27	F	8/4/14	10/2/14	59	
	GimXN inf-23	M	8/4/14	9/30/14	57	
	GimXN inf-	F	8/12/14	9/8/14	27	
	GimXN inf-8	F	8/12/14	9/22/14	41	
	GimXN inf-3	M	8/12/14	9/11/14	30	
	GimXN inf-4	M	8/12/14	9/15/14	34	
	GimXN inf-5	M	8/12/14	9/16/14	35	

	GimXN inf-6	M	8/12/14	9/16/14	35	
	GimXN inf-7	M	8/12/14	9/18/14	37	
	GimXN inf-9	M	8/12/14	9/22/14	41	
	GimXN inf-15	F	8/12/14	9/23/14	42	
	GimXN inf-24	F	8/12/14	9/30/14	49	
	GimXN inf-30	F	8/12/14	10/6/14	55	
	GimXN inf-35	F	8/12/14	10/21/14	70	
	GimXN inf-19	M	8/12/14	9/30/14	49	
	GimXN inf-34	M	8/12/14	10/20/14	69	
	GimXN inf-31	F	8/18/14	10/6/14	49	
	GimXN inf-37	F	8/18/14	10/27/14	70	
	GimXN inf-38	F	8/18/14	10/30/14	73	
	GimXN inf-40	F	8/18/14	11/4/14	78	
	GimXN inf-17	M	8/18/14	9/30/14	43	
	GimXN inf-14	F	8/18/14	9/23/14	36	
	GimXN inf-25	F	8/18/14	10/1/14	44	
	GimXN inf-18	F	8/18/14	9/30/14	43	
	GimXN inf-16	F	8/18/14	9/24/14	37	
	GimXN inf-12	M	8/18/14	9/23/14	36	
	GimXN inf-13	M	8/18/14	9/23/14	36	
	GimXN inf-21	M	8/18/14	9/30/14	43	
	GimXN inf-20	M	8/18/14	9/30/14	43	
	GimXN inf-10	M	8/18/14	9/23/14	36	
	GimXN inf-11	M	8/18/14	9/23/14	36	
	GimXN inf-26	M	8/18/14	10/1/14	44	
	GimXN inf-22	M	8/18/14	9/30/14	43	
	GimXN inf-28	M	8/18/14	10/6/14	49	
	GimXN inf-29	M	8/18/14	10/6/14	49	
	GimXN inf-39	F	9/1/14	10/31/14	60	
	GimXN inf-32	M	9/1/14	10/6/14	35	
	GimXN inf-33	M	9/1/14	10/7/14	36	
	GimXN inf-36	M	9/1/14	10/21/14	50	
	GimXN inf-41	M	9/1/14	11/13/14	73	
	GimXN inf-		9/1/14	9/10/14	9	thymus samples
	GimXN inf-		9/1/14	9/10/14	9	thymus samples
	GimXN inf-		9/1/14	9/10/14	9	thymus samples
uninfected	Gim-649	F	8/10/14	10/31/14	82	
	Gim-664	F	8/10/14	12/3/14	115	
	Gim-631	M	8/10/14	9/24/14	45	
	Gim-632	M	8/10/14	10/6/14	57	
	Gim-633	M	8/10/14	10/16/14	67	
	Gim-660	M	8/10/14	11/13/14	95	
	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-		9/16/14	9/26/14	10	thymus samples

	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-		9/16/14	9/26/14	10	thymus samples
	Gim-647	F	9/7/14	10/30/14	53	
	Gim-652	F	9/7/14	11/5/14	59	
	Gim-655	F	9/7/14	11/10/14	64	
	Gim-658	F	9/7/14	11/11/14	65	
	Gim-662	F	9/7/14	11/18/14	72	
	Gim-665	F	9/7/14	12/8/14	92	
	Gim-639	M	9/7/14	10/24/14	47	
	Gim-634	F	9/10/14	10/20/14	40	
	Gim-638	F	9/10/14	10/21/14	41	
	Gim-650	F	9/10/14	10/31/14	51	
	Gim-661	F	9/10/14	11/17/14	68	
	Gim-637	M	9/10/14	10/20/14	40	
	Gim-640	M	9/10/14	10/24/14	44	
	Gim-641	M	9/10/14	10/24/14	44	
	Gim-642	M	9/10/14	10/27/14	47	
	Gim-663	M	9/10/14	11/27/14	78	
	Gim-657	M	9/10/14	11/10/14	61	
	Gim-656	M	9/10/14	11/10/14	61	
	Gim-635	F	9/10/14	10/20/14	40	
	Gim-644	F	9/10/14	10/28/14	48	
	Gim-654	F	9/10/14	11/5/14	56	
	Gim-	M	9/10/14	10/27/14	47	found dead
	Gim-636	M	9/10/14	10/20/14	40	
	Gim-643	M	9/10/14	10/27/14	47	
	Gim-645	M	9/10/14	10/28/14	48	
	Gim-646	M	9/10/14	10/29/14	49	
	Gim-651	M	9/10/14	11/4/14	55	
	Gim-653	M	9/10/14	11/5/14	56	
	Gim-659	M	9/10/14	11/12/14	63	
	Gim-667	F	9/15/14	5/1/15	228	Outlier
	Gim-648	M	9/15/14	10/30/14	45	
	Gim-666	M	9/15/14	1/26/15	133	Outlier

Appendix 2. *MYC/Runx2* mice infected with WT MLV and MLV IN TP^{-} viruses

	ID	Sex	DoB	DoD	Life	reason for death
WT MLV	Gim16(WT) inf-9	F	2/16/16	4/8/16	52	
	Gim16(WT) inf-10	F	2/16/16	4/8/16	52	
	Gim16(WT) inf-13	F	2/16/16	4/19/16	63	
	Gim16(WT) inf-14	F	2/16/16	4/19/16	63	
	Gim16(WT) inf-11	M	2/16/16	4/8/16	52	
	Gim16(WT) inf-17	M	2/16/16	4/20/16	64	
	Gim16(WT) inf-1	F	3/1/16	4/4/16	34	
	Gim16(WT) inf-2	F	3/1/16	4/4/16	34	
	Gim16(WT) inf-3	F	3/1/16	4/4/16	34	
	Gim16(WT) inf-8	F	3/1/16	4/6/16	36	
	Gim16(WT) inf-	F	3/1/16	3/24/16	23	
	Gim16(WT) inf-4	M	3/1/16	4/4/16	34	
	Gim16(WT) inf-5	M	3/1/16	4/4/16	34	
	Gim16(WT) inf-6	M	3/1/16	4/4/16	34	
	Gim16(WT) inf-7	M	3/1/16	4/4/16	34	
	Gim16(WT) inf-15	F	3/9/16	4/19/16	41	
	Gim16(WT) inf-18	F	3/9/16	5/4/16	56	
	Gim16(WT) inf-19	F	3/9/16	5/4/16	56	
	Gim16(WT) inf-12	M	3/9/16	4/8/16	30	
	Gim16(WT) inf-16	M	3/9/16	4/19/16	41	
	Gim16(WT) inf-20	F	4/5/16	5/9/16	34	
	Gim16(WT) inf-21	F	4/5/16	5/9/16	34	
	Gim16(WT) inf-22	F	4/5/16	5/9/16	34	
	Gim16(WT) inf-23	F	4/5/16	5/9/16	34	
	Gim16(WT) inf-24	F	4/5/16	5/10/16	35	
	Gim16(WT) inf-25	F	4/5/16	5/10/16	35	
	Gim16(WT) inf-26	F	4/5/16	5/10/16	35	
	Gim16(WT) inf-28	F	4/5/16	5/16/16	41	
	Gim16(WT) inf-27	M	4/5/16	5/10/16	35	
	Gim16(WT)d10 inf-1		5/3/16	5/12/16	9	thymus samples
	Gim16(WT)d10 inf-2		5/3/16	5/12/16	9	thymus samples
	Gim16(WT)d10 inf-3		5/3/16	5/12/16	9	thymus samples
	Gim16(WT)d10 inf-4		5/3/16	5/12/16	9	thymus samples
	Gim16(WT)d10 inf-5		5/3/16	5/12/16	9	thymus samples
	Gim16(WT) inf-		5/3/16	5/18/16	15	excess
	Gim16(WT) inf-		5/3/16	5/18/16	15	excess
	Gim16(WT) inf-		5/3/16	5/18/16	15	excess

	Gim16(WT) inf-		5/3/16	5/18/16	15	excess
	Gim16(WT) inf-		5/3/16	5/18/16	15	excess
MLV IN TP^-	Gim16(TP-)inf-3	F	2/16/16	4/4/16	48	
	Gim16(TP-)inf-6	F	2/16/16	4/13/16	57	
	Gim16(TP-)inf-9	F	2/16/16	4/19/16	63	
	Gim16(TP-)inf-11	F	2/16/16	4/19/16	63	
	Gim16(TP-)inf-1	M	2/16/16	4/4/16	48	
	Gim16(TP-)inf-2	M	2/16/16	4/4/16	48	
	Gim16(TP-)inf-10	M	2/16/16	4/19/16	63	
	Gim16(TP-)inf-4	M	2/16/16	4/6/16	50	
	Gim16(TP-)inf-5	M	2/16/16	4/12/16	56	
	Gim16(TP-)inf-8	M	2/16/16	4/15/16	59	
	Gim16(TP-)inf-14	M	2/16/16	4/25/16	69	
	Gim16(TP-)inf-		2/16/16	NA	NA	lost
	Gim16(TP-)inf-		2/16/16	NA	NA	lost
	Gim16(TP-)inf-		2/16/16	NA	NA	lost
	Gim16(TP-)inf-7	F	3/7/16	4/13/16	37	
	Gim16(TP-)inf-21	F	3/7/16	5/16/16	70	
	Gim16(TP-)inf-20	M	3/7/16	5/11/16	65	
	Gim16(TP-)inf-22	M	3/7/16	5/16/16	70	
	Gim16(TP-)inf-23	M	3/7/16	5/18/16	72	
	Gim16(TP-)inf-16	F	3/23/16	4/26/16	34	
	Gim16(TP-)inf-12	M	3/23/16	4/20/16	28	
	Gim16(TP-)inf-13	M	3/23/16	4/22/16	30	
	Gim16(TP-)inf-15	M	3/23/16	4/25/16	33	
	Gim16(TP-)inf-17	M	3/23/16	4/26/16	34	
	Gim16(TP-)inf-18	M	3/23/16	4/28/16	36	
	Gim16(TP-)inf-19	M	3/23/16	4/28/16	36	
	Gim16(TP-)d10 inf-1		5/3/16	5/12/16	9	thymus samples
	Gim16(TP-)d10 inf-2		5/3/16	5/12/16	9	thymus samples
	Gim16(TP-)d10 inf-3		5/3/16	5/12/16	9	thymus samples
	Gim16(TP-)d10 inf-4		5/3/16	5/12/16	9	thymus samples
	Gim16(TP-)d10 inf-5		5/3/16	5/12/16	9	thymus samples
	Gim16(TP-)inf-		5/3/16	5/18/16	15	excess
	Gim16(TP-)inf-		5/3/16	5/18/16	15	excess
	Gim16(TP-)inf-		5/3/16	5/18/16	15	excess

* 🟦 blue box- tumors from these mice were analyzed by next-generation sequencing

Appendix 3. Genomic annotations and ChIPSeq datasets used in the study

Dataset	Accession number	Genome (reference build)
Brd4	GSM937540	Mouse (mm8)*
H3K27ac	ENCFF001KYC	Mouse (mm10)
H3K4me1	GSM1000102	Mouse (mm9)*
H3K4me3	GSM1000101	Mouse (mm9)*
ChromHMM definition states	wgEncodeBroadHmmK562HMM	Human (hg18)**
H3K27me3	GSM733658	Human (hg19)
H3K9me3	GSM733776	Human (hg19)
[a]9 state hi H3K27ac	ENCSR032YTK	Human (hg19)
H3K36me3	GSM733714	Human (hg19)
H2A.Z	GSM733786	Human (hg19)
H3K9me1	GSM733777	Human (hg19)
H3K9me3	GSM733776	Human (hg19)
H3K4me1	GSM733692	Human (hg19)
H3K79me2	GSM733653	Human (hg19)
H3K4me2	GSM733651	Human (hg19)
H3K4me3	GSM733680	Human (hg19)
Brd4	GSM2635249	Human (hg19)
CTCF	GSM733719	Human (hg19)
H3K9ac	GSM733778	Human (hg19)

aDataset was obtained from ENCODE, whereas other datasets were from NCBI Gene Expression Omnibus.
*Genomic coordinates were lifted over from mm8 to mm10.
**Genomic coordinates were lifted over from hg18 to hg19.

Appendix 4. Fisher's test for statistical comparison of integration profile

Mouse tumors from *MYC/Runx2* mice model*

		NC	WT6	WT8	WT10	WT12	TP^{-}4	TP^{-} 6	TP^{-}7	TP^{-}9	TP^{-}16
TSS	NC	1	2.50E-65	1.38E-44	9.02E-47	1.14E-32	3.66E-47	9.42E-32	1.49E-46	1.48E-25	2.51E-08
	WT6										4.39E-06
CpG		NC									
	NC	1	4.44E-61	6.23E-40	1.00E-43	2.12E-38	3.66E-47	9.42E-32	1.49E-46	1.48E-25	3.66E-47
	WT6										**0.00013074**

K562 infected cells**

		RIC	WT	IN TP^{-}
TSS	RIC	1	< 2.2 x10-299	7.42E-20
	WT			5.65E-44
CpG		RIC	WT	IN TP^{-}
	RIC	1	< 2.2 x10-299	1.16E-25
	WT			3.13E-47

K562 infected cells (15-chromatin states)**

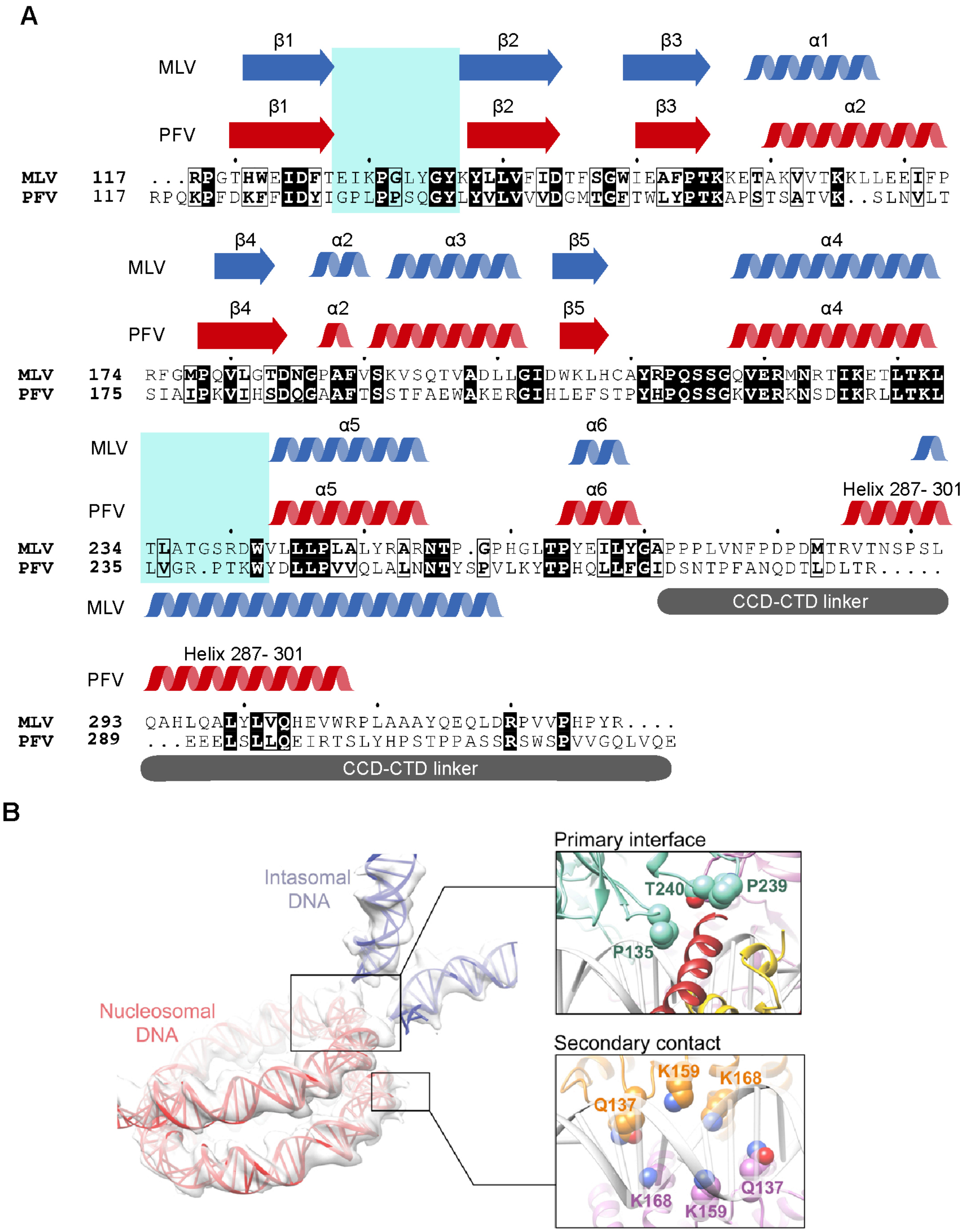
		RIC	WT	IN TP^{-}
State 11	RIC	1	6.63E-40	4.03E-15
	WT			2.08E-50
State 13		RIC	WT	IN TP^{-}
	RIC	1	< 2.2 x10-299	1.71E-119
	WT			2.92E-61

*P-values are shown for NC vs. tumor samples, tumor WT6 vs. TP^{-}16
**P-values for infected K562 samples are shown for RIC vs. infected K562 cells and WT vs IN TP^{-}

Appendix 5. Chromatin-binding peptides inserted into MLV IN XN

Peptides		
HP1	Heterchromatin protein 1	NKGAKPVVVLQKLS
TIF1β	Transcription intermediary factor	GLLRKVPRVSLERLDLDLTSDSQPPVFK
PFV CBS	Prototype foamy virus chromatin binding sequence	QGGYNLRPRTYQP
KSHV LANA$_{1-23}$	Kaposi sarcoma herpesvirus latency-associated nuclear antigen	MAPPGMRLRSGRSTGAPLTRGSC

Appendix 6


A
β1
β2
β3
α1
MLV
β1
β2
β3
α2
PFV
MLV 117 ...RPGTHWEIDFTEIKPGLYGYKYLLVFIDTFSGWIEAFPTKKETAKVVTKKLLEEIFP
PFV 117 RPQKPFDKFFIDYIGPLPPSQGYLYVLVVVDGMTGFTWLYPTKAPSTSATVK..SLNVLT
β4
α2
α3
β5
α4
MLV
β4
α2
β5
α4
PFV
MLV 174 RFGMPQVLGTDNGPAFVSKVSQTVADLLGIDWKLHCAYRPQSSGQVERMNRTIKETLTKL
PFV 175 SIAIPKVIHSDQGAAFTSSTFAEWAKERGIHLEFSTPYHPQSSGKVERKNSDIKRLLTKL
MLV
α5
α6
α5
α6
Helix 287- 301
PFV
MLV 234 TLATGSRDWVLLLPLALYRARNTP.GPHGLTPYEILYGAPPPLVNFPDPDMTRVTNSPSL
PFV 235 LVGR.PTKWYDLLPVVQLALNNTYSPVLKYTPHQLLFGIDSNTPFANQDTLDLTR.....
MLV
CCD-CTD linker
Helix 287- 301
PFV
MLV 293 QAHLQALYLVQHEVWRPLAAAYQEQLDRPVVPHPYR....
PFV 289 ...EEELSLLQEIRTSLYHPSTPPASSRSWSPVVGQLVQE
CCD-CTD linker
B
Intasomal DNA
Nucleosomal DNA
Primary interface
T240
P239
P135
Secondary contact
K159
K168
Q137
K168
K159
Q137

Sequence alignment of MLV IN CCD and PFV IN CCD. (A) Structure based sequence alignment of PFV IN CCD (117-325) and MLV IN CCD (117-328) with secondary structure prediction using PROMALS3D (155) and displayed using ESPript (156). PFV secondary structures (red helices) are derived from PFV intasome structure (3OS1). MLV secondary structures (blue) are based on the homology model (59). Sequences in aquamarine boxes are loop regions where PFV IN intasome are predicted to interact with nucleosomal proteins. (B) This figure is adapted from (64). The cryo-EM structure of PFV intasome bound to nucleosome reveal residues (aquamarine) that interact with histone H2A protein (red).